US006467023B1

(12) **United States Patent**
DeKoning et al.

(10) **Patent No.: US 6,467,023 B1**
(45) **Date of Patent: Oct. 15, 2002**

(54) **METHOD FOR LOGICAL UNIT CREATION WITH IMMEDIATE AVAILABILITY IN A RAID STORAGE ENVIRONMENT**

(75) Inventors: **Rodney A. DeKoning**, August, KS (US); **Donald R. Humlicek**, Wichita, KS (US); **Robin Huber**, Wichita, KS (US)

(73) Assignee: **LSI Logic Corporation**, Milpitas, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/274,582**

(22) Filed: **Mar. 23, 1999**

(51) **Int. Cl.**[7] **G06F 12/02**
(52) **U.S. Cl.** **711/114**; 714/6
(58) **Field of Search** 711/114, 168, 711/170; 714/6

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,390,327	A	*	2/1995	Lubbers et al. 714/7
5,394,532	A	*	2/1995	Belsan 395/425
5,430,855	A	*	7/1995	Walsh et al. 395/275
5,450,384	A	*	9/1995	Dahman et al. 369/30
5,522,031	A	*	5/1996	Ellis et al. 714/6
5,546,558	A	*	8/1996	Jacobson et al. 395/441
5,574,851	A	*	11/1996	Rathunde 711/114
5,621,882	A	*	4/1997	Kakuta 395/482.04
5,657,468	A	*	8/1997	Stallmo et al. 395/441
5,706,472	A	*	1/1998	Ruff et al. 395/497.04
5,708,769	A	*	1/1998	Stallmo 395/182.04
5,812,753	A	*	9/1998	Chiariotti 714/6
6,070,170	A	*	5/2000	Friske et al. 707/200

OTHER PUBLICATIONS

American Megatrends RAID Overview; AST Computer; http://www.ejs.is/voruuppl/ast/ami-raid.htm: pp. 1–25.*

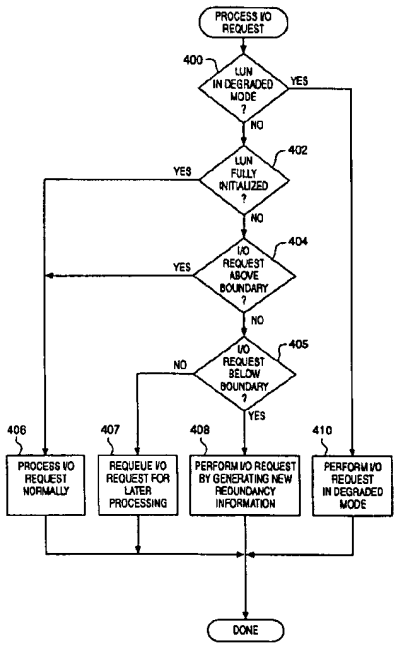
* cited by examiner

Primary Examiner—Hiep T. Nguyen
(74) *Attorney, Agent, or Firm*—Lathrop & Gage

(57) **ABSTRACT**

Methods and associated structure for enabling immediate availability of a disk array storage device. In particular, the methods and associated structure of the present invention permit access to a logical unit of a storage system immediately following creation of the logical unit. Initialization of the logical unit to initialize redundancy information therein proceeds in parallel with host system access to the storage space of the logical unit. The initialization process maintains a boundary parameter value indicative of the progress of the initialization process. Storage space above the boundary has had its redundancy information initialized while storage space below the boundary has not. Where an I/O request is entirely above the boundary, it is processed normally in accordance with the management of the logical unit. Where part of an I/O request is below the boundary, it is processed in a special manner that assures integrity of the redundancy data.

14 Claims, 6 Drawing Sheets



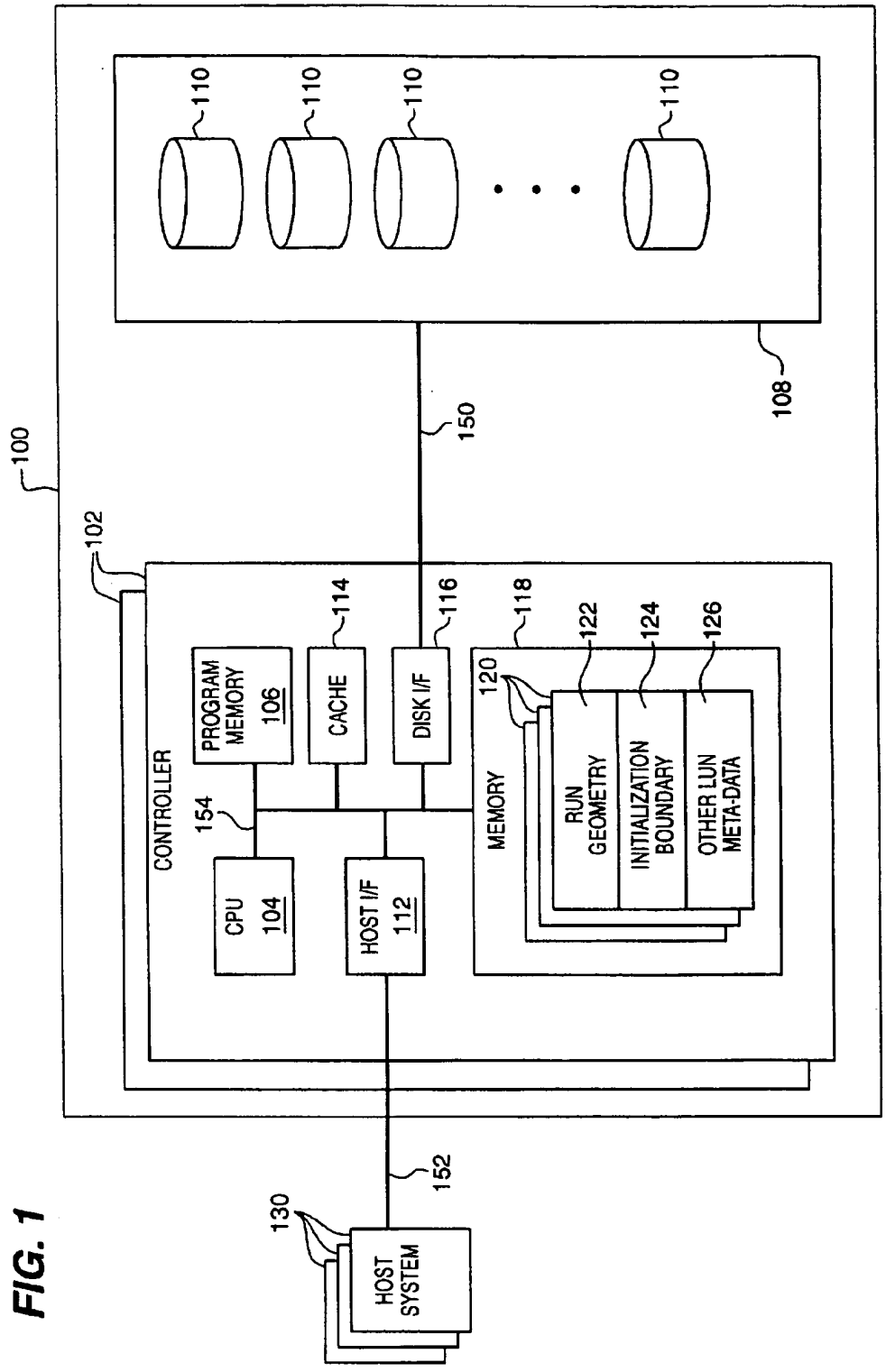

100
102
108
110
110
110
110
150
152
130
HOST SYSTEM
CONTROLLER
154
CPU 104
HOST I/F 112
PROGRAM MEMORY 106
CACHE
DISK I/F
114
116
118
120
122
124
126
MEMORY
RUN GEOMETRY
INITIALIZATION BOUNDARY
OTHER LUN META-DATA

FIG. 1

FIG. 2

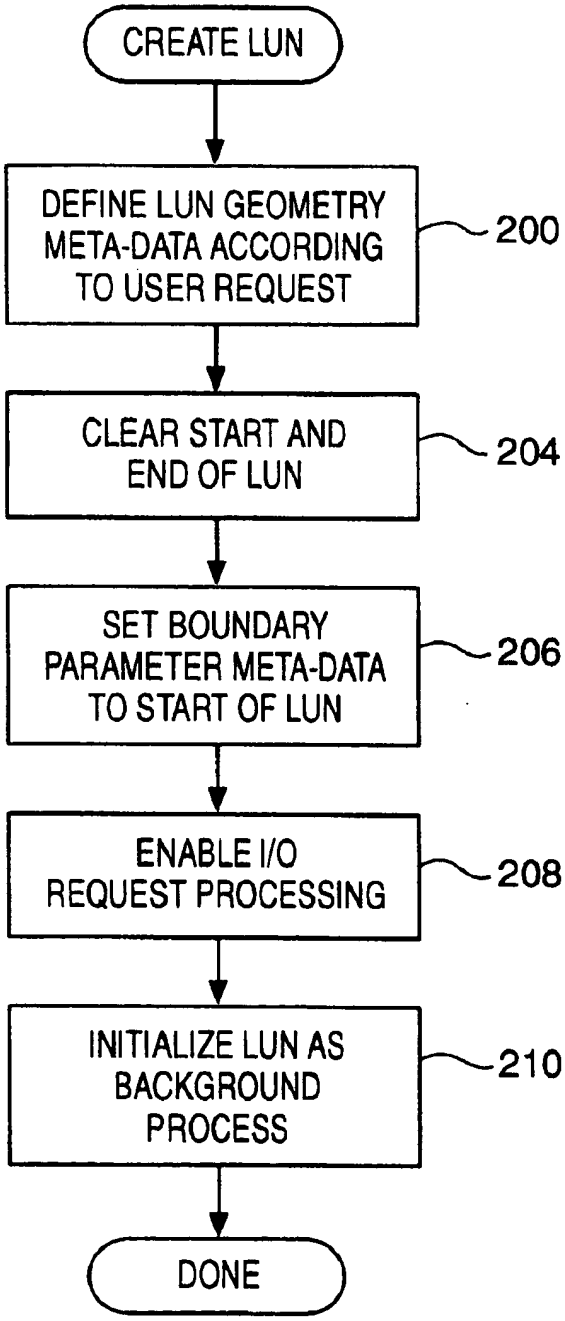

CREATE LUN
DEFINE LUN GEOMETRY META-DATA ACCORDING TO USER REQUEST
200
CLEAR START AND END OF LUN
204
SET BOUNDARY PARAMETER META-DATA TO START OF LUN
206
ENABLE I/O REQUEST PROCESSING
208
INITIALIZE LUN AS BACKGROUND PROCESS
210
DONE

FIG. 3

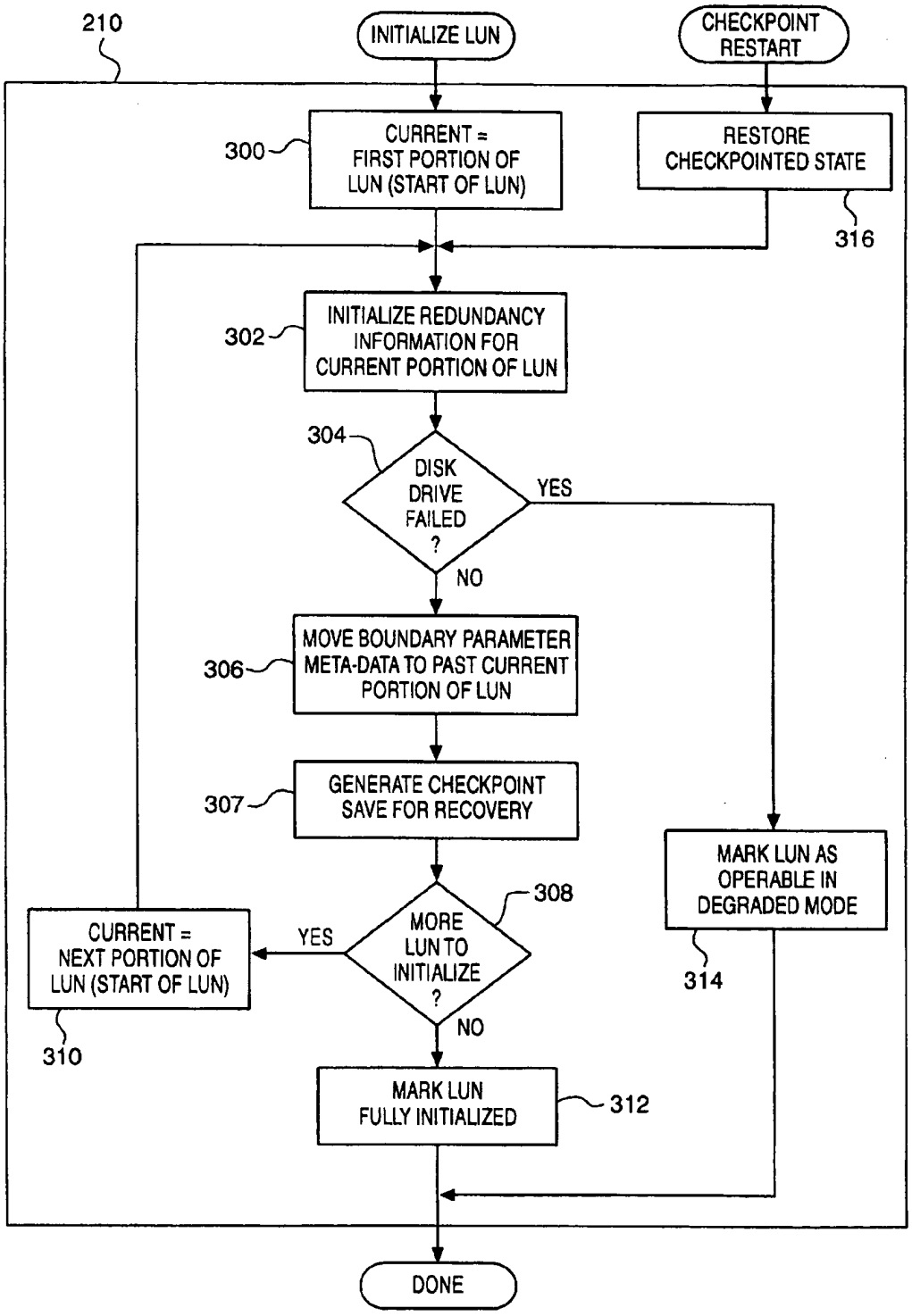

210
INITIALIZE LUN
CHECKPOINT RESTART
300
CURRENT = FIRST PORTION OF LUN (START OF LUN)
RESTORE CHECKPOINTED STATE
316
302
INITIALIZE REDUNDANCY INFORMATION FOR CURRENT PORTION OF LUN
304
DISK DRIVE FAILED ?
YES
NO
306
MOVE BOUNDARY PARAMETER META-DATA TO PAST CURRENT PORTION OF LUN
307
GENERATE CHECKPOINT SAVE FOR RECOVERY
MARK LUN AS OPERABLE IN DEGRADED MODE
314
308
MORE LUN TO INITIALIZE ?
YES
CURRENT = NEXT PORTION OF LUN (START OF LUN)
310
NO
MARK LUN FULLY INITIALIZED
312
DONE

FIG. 4

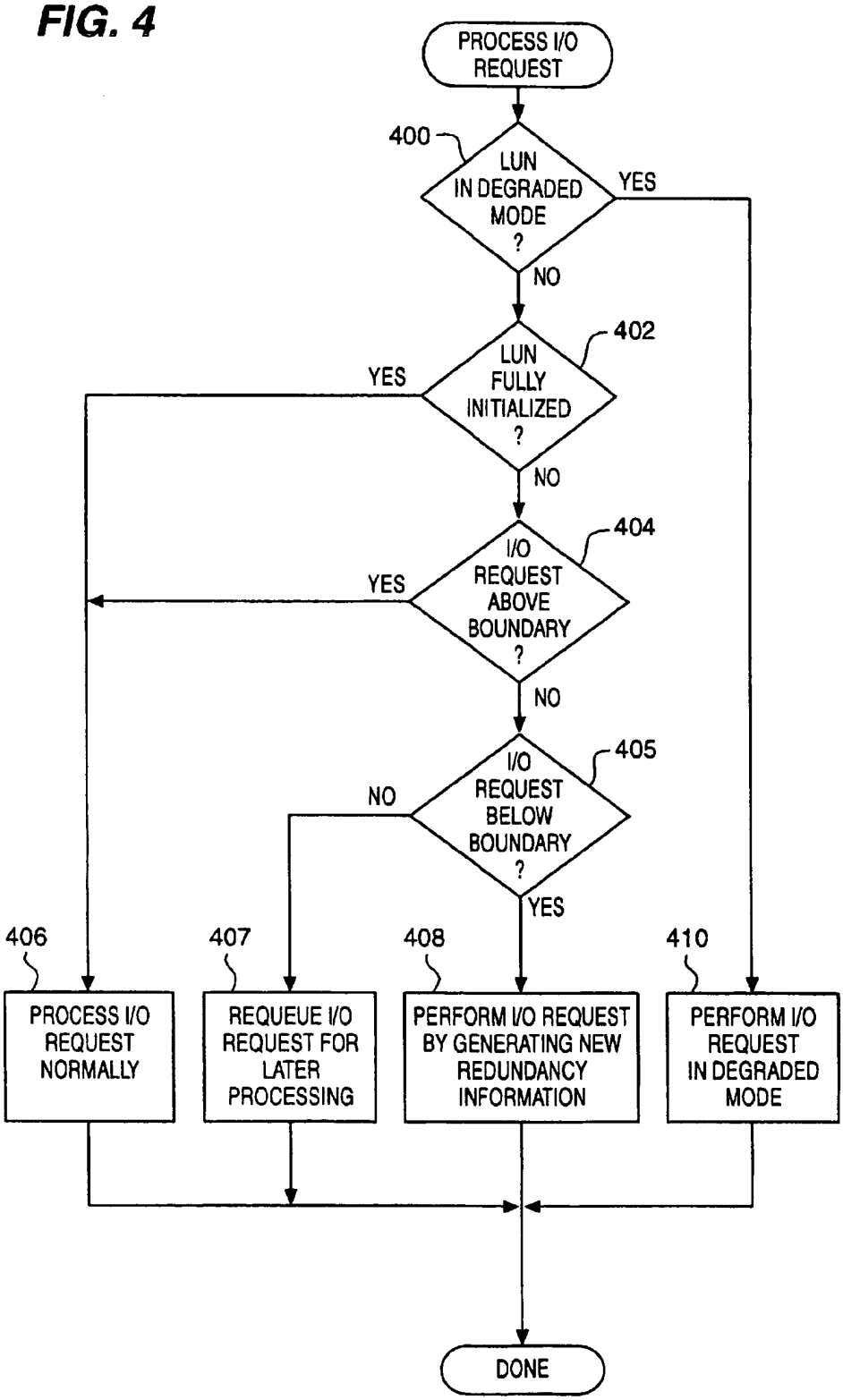

PROCESS I/O REQUEST
400
LUN IN DEGRADED MODE ?
YES
NO
402
LUN FULLY INITIALIZED ?
YES
NO
404
I/O REQUEST ABOVE BOUNDARY ?
YES
NO
405
I/O REQUEST BELOW BOUNDARY ?
NO
YES
406
PROCESS I/O REQUEST NORMALLY
407
REQUEUE I/O REQUEST FOR LATER PROCESSING
408
PERFORM I/O REQUEST BY GENERATING NEW REDUNDANCY INFORMATION
410
PERFORM I/O REQUEST IN DEGRADED MODE
DONE

FIG. 5

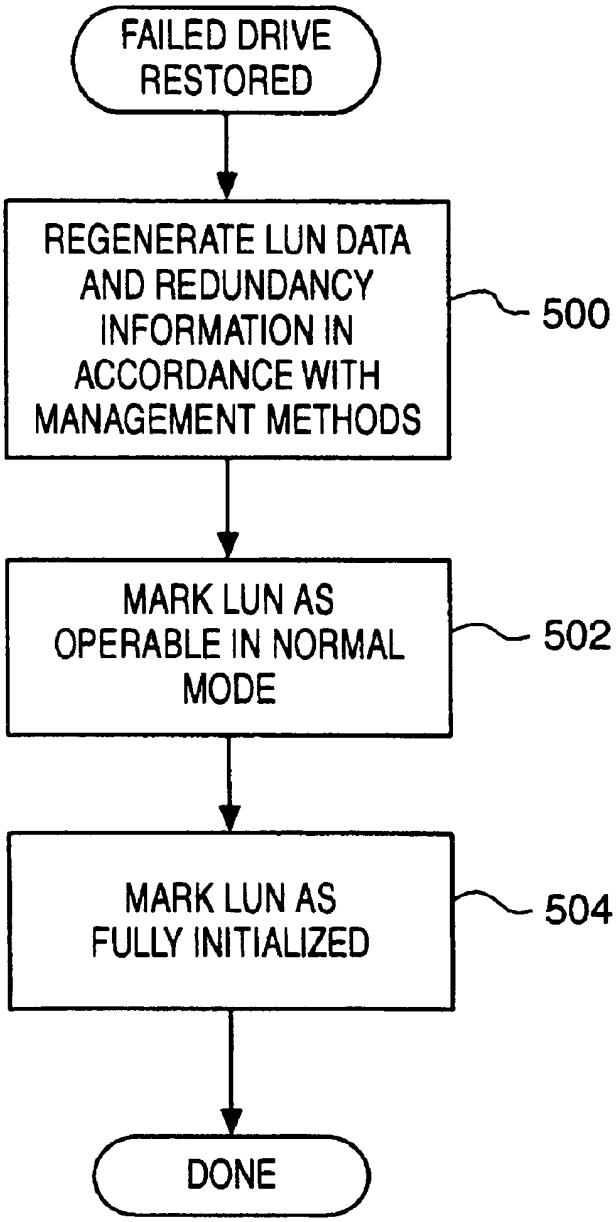
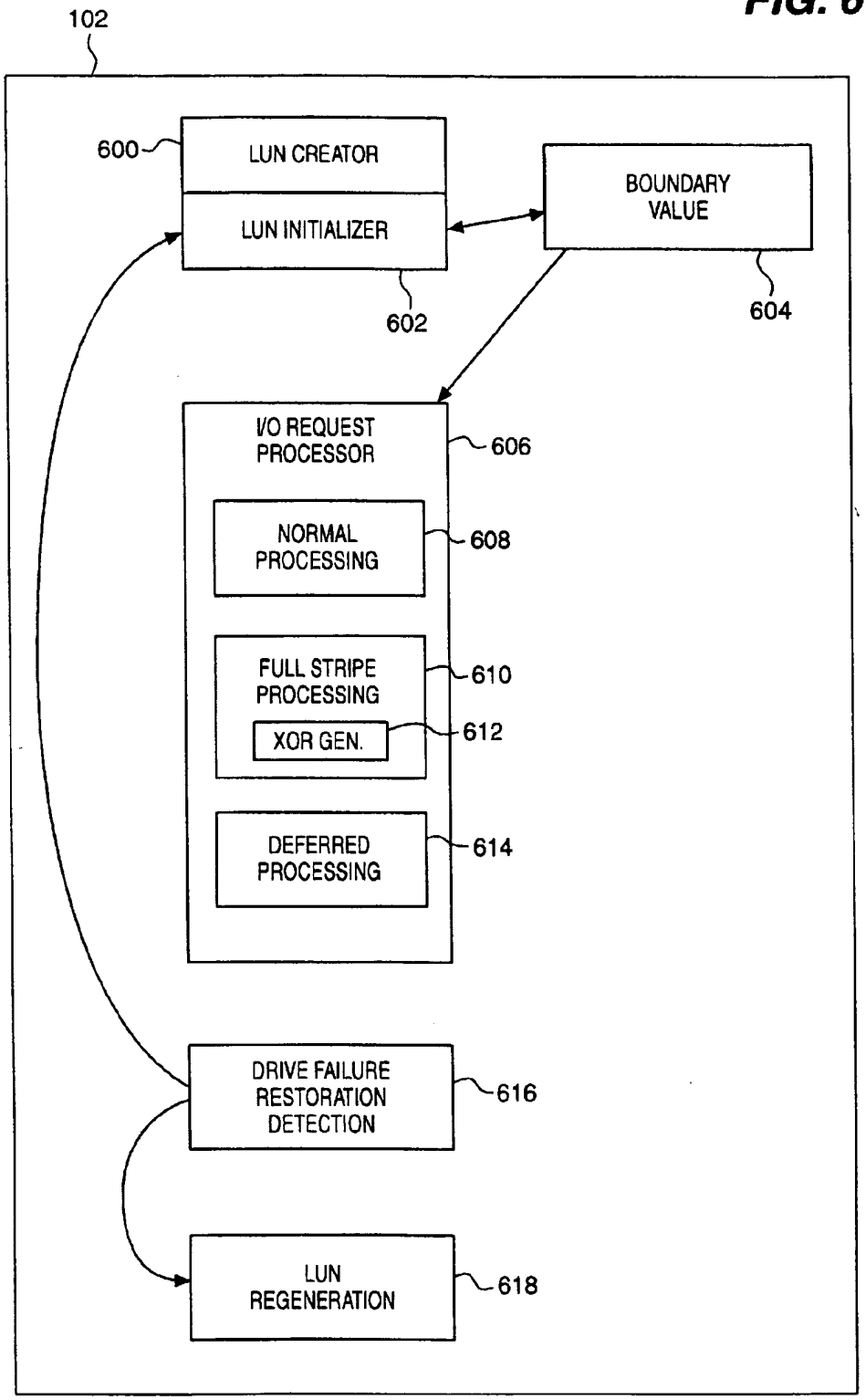

FAILED DRIVE RESTORED
REGENERATE LUN DATA AND REDUNDANCY INFORMATION IN ACCORDANCE WITH MANAGEMENT METHODS
500
MARK LUN AS OPERABLE IN NORMAL MODE
502
MARK LUN AS FULLY INITIALIZED
504
DONE

FIG. 6


102
600
LUN CREATOR
LUN INITIALIZER
602
BOUNDARY VALUE
604
I/O REQUEST PROCESSOR
606
NORMAL PROCESSING
608
FULL STRIPE PROCESSING
610
XOR GEN.
612
DEFERRED PROCESSING
614
DRIVE FAILURE RESTORATION DETECTION
616
LUN REGENERATION
618

METHOD FOR LOGICAL UNIT CREATION WITH IMMEDIATE AVAILABILITY IN A RAID STORAGE ENVIRONMENT

BACKGROUND OF THE INVENTION

1. Field of the Invention

The invention relates to a method for improving access to a newly defined logical unit in a storage system and in particular relates to methods and structures for creating a RAID logical unit so as to make it available immediately for processing of host I/O requests.

2. Description of Related Art

Typical computing systems store data within devices such as hard disk drives, floppy drives, tape, compact disk, etc. These devices are otherwise known as storage devices. The storage capacity of these storage devices has rapidly increased as computing applications' demand for storage have increased. Simultaneous with the increase in capacity, computer applications and users have demanded increased performance. Computing applications have become highly dependent on high performance, high capacity storage devices. However, such increased dependency on storage devices underscores the need for increased reliability of such storage subsystems. Failure of such high capacity, high performance storage devices and subsystems can cripple vital computing applications.

Disk array storage systems provide both improved capacity and performance as compared to single disk devices. In a disk array storage system, a plurality of disk drives are used in a cooperative manner such that multiple disk drives are performing, in parallel, the tasks normally performed by a single disk drive. Striping techniques are often used to spread large amounts of information over a plurality of disk drives in a disk array storage system. So spreading the data over multiple disk drives improves perceived performance of the storage system in that a large I/O operation is processed by multiple disk drives in parallel rather than being queued awaiting processing by a single disk drive.

However, adding multiple disk drives to a storage system reduces to reliability of the overall storage system. In particular, spreading data over multiple disk drives in a disk array increases the potential for system failure. Failure of any of the multiple disk drives translates to failure of the storage system because the data stored thereon cannot be correctly retrieved.

RAID techniques are commonly used to improve reliability in disk array storage systems. RAID techniques generally configure multiple disk drives in a disk array in geometries that permit redundancy of stored data to assure data integrity in case of various failures. In many such redundant subsystems, recovery from many common failures can be automated within the storage subsystem itself due to the use of data redundancy, error codes, and so-called "hot spares" (extra disk drives that may be activated to replace a failed, previously active disk drive). The 1987 publication by David A. Patterson, et al., from University of California at Berkeley entitled *A Case for Redundant Arrays of Inexpensive Disks(RAID)*, reviews the fundamental concepts of RAID technology.

RAID level zero, also commonly referred to as disk striping, distributes data As stored on a storage subsystem across a plurality of disk drives to permit parallel operation of a plurality of disk drives thereby improving the performance of I/O write requests to the storage subsystem.

Though RAID level zero functionality improves I/O write operation performance, reliability of the disk array subsystem is decreased as compared to that of a single large disk drive. To improve reliability of disk arrays, other RAID geometries for data storage include generation and storage of redundancy information to permit continued operation of the disk array through certain common failure modes of the disk drives in the disk array.

There are six other "levels" of standard RAID geometries that include redundancy information as defined in the Patterson publication. Other RAID geometries have been more recently adopted and utilize similar concepts. For example, RAID level six provides additional redundancy to enable continued operation even in the case of failure of two disk drives in a disk array.

The simplest array, a RAID level 1 system, comprises one or more disks for storing data and an equal number of additional "mirror" disks for storing copies of the information written to the data disks. The remaining RAID levels, identified as RAID levels 2, 3, 4 and 5 systems by Patterson, segment the data into portions for storage across several data disks. One or more additional disks are utilized to store error check or parity information. RAID level 6 further enhances reliability by adding additional redundancy information to permit continued operation through multiple disk failures. The methods of the present invention may be useful in conjunction with any of the standard RAID levels.

RAID storage subsystems typically utilize a control module (controller) that shields the user or host system from the details of managing the redundant array. The controller makes the subsystem appear to the host computer as one (or more), highly reliable, high capacity disk drive. In fact, the RAID controller may distribute the host computer system supplied data across a plurality of the small independent drives with redundancy and error checking information so as to improve subsystem reliability. The mapping of a logical location of the host supplied data to a physical location on the array of disk drives is performed by the controller in a manner that is transparent to the host system. RAID level 0 striping for example is transparent to the host system. The data is simply distributed by the controller over a plurality of disks in the disk array to improve overall system performance.

RAID storage systems generally subdivide the disk array storage capacity into distinct partitions referred to as logical units (LUNs). Each logical unit may be managed in accordance with a selected RAID management technique. In other words, each LUN may use a different RAID management level as required for its particular application.

A typical sequence in configuring LUNs in a RAID system involves a user (typically a system administrator) defining storage space to create a particular LUN. With the storage space so defined, a preferred RAID storage management technique is associated with the newly created LUN. The storage space of the LUN is then typically initialized—a process that involves formatting the storage space associated with the LUN to clear any previously stored data and involves initializing any redundancy information required by the associated RAID management level.

The formatting and other initialization a LUN is generally a time-consuming process. Formatting all storage space for a large LUN may take several minutes or even hours. Often, the initialization processing further includes a verification step to verify proper access to all storage space within the LUN. Such verification may include writing and reading of test data or commands on each disk of the LUN to verify

proper operation of the LUN. Errors detected in this verification process that relate to recording media defects are then repaired if possible (i.e., by allocating a spare sector or track of the associated disk drive to replace a defective one). This verification process is yet another time consuming aspect of the LUN initialization process.

Most presently known RAID storage systems preclude processing of host I/O requests for the logical unit until after the initialization has completed. In such systems, the logical unit is said to be unavailable until initialization has completed. Some RAID storage systems have improved upon these known systems by making the logical unit available as soon configuration information has been saved to define the mapping of data within the logical unit and as soon redundancy information has been initialized for the entire LUN. Time consuming disk media verification is skipped in lieu of sensing and correcting such errors as the LUN is used.

It remains a problem however unit in such improved systems, to make the logical unit available as soon as possible for processing of host I/O requests. Initialization of redundancy information still requires significant amount of time wherein large logical unit is defined. Further, presently known logical unit initialization sequences cease operation of the LUN when one of the drives in the logical unit fails. Further access to the logical unit having failed drives is then precluded in presently known RAID systems.

It is therefore the problem to make a newly defined RAID logical unit available for host I/O processing as soon as possible, preferably even during initialization. It is a further problem to assure that the logical unit is available for processing of host I/O requests unit when a drive in the LUN is sensed to have failed during the initialization sequence.

SUMMARY OF THE INVENTION

The present invention solves the above and other problems, thereby advancing the state of the useful arts, by providing methods and associated structure to permit immediate availability of newly defined LUNs as soon as possible after configuration information regarding the LUN has been saved. The present invention allows I/O requests to be performed on a logical unit in parallel with initialization of the redundancy information in the LUN. Furthermore, the present invention assures continued availability of the logical unit and eventual completion of LUN initialization despite failure of a disk drive during the initialization process. Methods of the present invention permit such a LUN with a failed drive to continue operation though in a degraded mode.

More specifically, methods and structures of the present invention permit immediate availability of a newly created logical unit for processing host I/O requests while redundancy information for the LUN is being initialized. Methods of the present invention maintain boundary information regarding progress in initializing redundancy information for the LUN. Host I/O requests for writing of data below this boundary level (i.e., in an area of the LUN for which redundancy information has not yet been initialized) are performed in a manner that assures valid redundancy data is generated and written. Conversely, host I/O requests for writing of data above the boundary level (i.e., in an area of the LUN for which valid redundancy information has already been initialized) proceed in a normal manner in accordance with the associated RAID level for the LUN.

In a further aspect of the present invention, initialization of redundancy information in accordance with the present invention as described above is terminated in response to detection of a failed disk drive in the LUN. However, unlike prior techniques, the LUN remains available for processing of HOST I/O requests though in a degraded mode. When the failed disk drive is eventually replaced, standard RAID regeneration techniques to recover loss data are used to complete the initialization of redundancy information for the LUN.

Another aspect of the invention lies in the capability to resume the initialization process. The present status of the initialization is maintained as each portion of the LUN is initialized. As the boundary separating initialized from un-initialized areas of the LUN is moved, checkpoint data is saved to enable recovery at any point in the initialization. If the initialization is interrupted, such as due to a power loss, the checkpoint information is used to resume the initialization where it left off.

It is therefore an object of the present invention to provide methods and associated structure that permit immediate availability of a newly created RAID logical unit.

It is another object of the present invention to provide methods and associated structure that permit immediate availability of a newly created RAID logical unit prior to initialization of all redundancy information in the logical unit.

It is a further object of the present invention to provide methods and associated structure that permit completion of initialization of a newly created LUN despite failure of a disk drive in the LUN.

It is still a further object of the present invention to provide methods and associated structures to permit interruption of the initialization of a newly created LUN such that the initialization is completed when a failed disk drive is replaced.

It is still another object of the present invention to provide methods and associated structure that permits host I/O requests to be processed in parallel with initialization of redundancy information in a newly defined logical unit.

It is yet another aspect of the present invention to provide methods and associated structure that permit completion of redundancy information initialization in a newly defined logical unit by use of standard RAID regeneration techniques in accordance with the RAID management level of the logical unit.

It is still another aspect of the present invention to provide methods and associated structure to permit interruption of the initialization procedure for a newly defined LUN such that the initialization may be resumed at a later time.

The above and other objects, aspects, features, and advantages of the present invention will become apparent from the following detailed description and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a block diagram of a storage system embodying the methods and structures in accordance with the present invention.

FIG. **2** is a flowchart describing a method of the present invention for creating a new logical unit in a storage system.

FIG. **3** is a flowchart describing a method of the present invention to initialize a newly created logical unit in parallel with processing of I/O requests directed to the logical unit.

FIG. **4** is a flowchart describing a method of the present invention to process I/O requests directed to a logical unit in parallel with initializing the newly created logical unit.

FIG. **5** is a flowchart describing a method of the present invention operable to complete initialization of a newly

created logical unit the initialization of which was prematurely terminated by sensing a drive failure.

FIG. **6** is a block diagram of the significant functional elements operable within the storage controller in accordance with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

While the invention is susceptible to various modifications and alternative forms, a specific embodiment thereof has been shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that it is not intended to limit the invention to the particular form disclosed, but on the contrary, the invention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the invention as defined by the appended claims.

Storage System of the Present Invention

FIG. 1 is a block diagram depicting a typical RAID storage system **100** embodying the methods and associated structures of the present invention. One or more controllers **102** receive and process I/O requests from one or more host systems **130** via path **152**. Processing of the I/O requests generally involves performing lower-level I/O operations on disk array **108** via path **150**. Disk array **108** comprises a plurality of disk drives **110** that are in turn logically subdivided by storage management methods of the controllers **102** into plurality of logical units (LUNs). Each LUN is managed by controller **102** in accordance with an associated RAID storage management level (i.e., RAID levels 0 through 6). Those skilled in the art will readily recognized that paths **152** and **150** may be implemented using any of several well-known industry standard communication media and protocols including, for example, SCSI, Fibre Channel, SSA, IPI, etc.

Each controller **102** generally comprises a CPU **104** that manages the overall operation of the controller **102** to process received I/O requests from host systems **130**. CPU **104** fetches program instructions from program memory **106** via processor bus **154**. Methods of the present invention are therefore preferably embodied as programmed instructions within program memory **106** operable within CPU **104** of controller **102**.

Host interface **112** provides the interface between CPU **104** and host systems **130** via path **152** and processor bus **154**. In like manner, disk interface **116** provides the interface between CPU **104** and disk array **108** via path **150** and processor bus **154**. As noted above, interfaces **112** and **116** are well-known commercially available devices designed to adapt signals between processor bus and either path **150** or **152** in accordance with the selected media and protocol of their respective interfaces.

Cache memory **114** enhances performance of controller **102** in processing I/O requests received from host systems **130**. In general, I/O write requests are processed by temporarily storing modified information in cache memory **114** and later flushing or posting such cached information through disk interface **116** to disk array **108** for persistent storage on disks **110**. In like manner, I/O read requests are more rapidly processed by retrieving requested information from cache memory **114**. Cache memory management methods and architectures are well-known to those skilled in the art.

Memory **118** is a random access memory subsystem utilized for purposes of storing and retrieving management information regarding each LUN in disk array **108**. Each LUN in disk array **108** has a corresponding LUN meta-data structure **120** that describes required management information to manipulate a corresponding LUN. Each LUN meta-data structure **120** includes the LUN's geometry information **122**, initialization boundary parameter value **124**, and other LUN meta-data **126**.

LUN geometry information **122** includes information required for RAID storage management of the LUN. Most RAID storage systems provide access to a LUN in storage units referred to as blocks. The user views the LUN as a sequence of logical blocks identified by sequential block numbers through the number of blocks allocated in the LUN. Depending on the RAID management level associated with the LUN, these logical block numbers are mapped or translated into corresponding physical block numbers—physical locations on the disk drives. LUN geometry information **122** therefore includes mapping information describing a logical to physical mapping of blocks within the LUNs as well as information regarding the desired RAID management level to be used in manipulating the corresponding LUNs. Such mapping of blocks in accordance with RAID management levels is well known to those skilled in the art.

Initialization boundary parameter value **124** is used, as described further herein below, to indicate the degree of progress in the initialization of a newly created LUN. Other status and configuration information regarding the particular LUNs is stored in other LUN meta-data **126**.

Those skilled in the art will readily recognized that FIG. 1 is intended to be representative of a wide variety of possible storage system architectures that may embody the methods and structures of the present invention. FIG. **1** therefore represents all such storage systems and controller architectures in which the methods and associated structures of the present invention may be advantageously applied.

LUN Creation Method

FIG. **2** is a flowchart describing a method of the present invention for creating a newly defined LUN within the storage subsystem. Element **200** in his first operable to define LUN geometry meta-data information according to the user's request for creating a new logical unit. As noted above, the geometry meta-data describes, among other things, the logical to physical mapping of data in the LUN in accordance with the user's selected RAID management level and other user defined parameters.

Element **204** is next operable to clear a small portion of the storage space of the LUN. The storage areas near the start and end of a LUN are often utilized for file system information by an operating system. Clearing small portions at the beginning and end of the LUN may help such file system software to identify the LUN as un-initialized. This heuristic approach as helpful to a number of common storage applications such as file system applications in common, commercially available operating systems. Those skilled in the art recognized that the methods of the present invention largely rely on system and application software that is not dependent on the storage space all being initialized. Element **204** is therefore a heuristic attempt to prevent many common application and system programs from viewing the LUN as initialized with "random" data. The processing of element **204** is not however a necessary step for application of the methods of the present invention. A large class of storage applications do not rely on the storage space of a LUN to be cleared or initialized prior to use.

Element **206** is next operable to set the boundary parameter value in the meta-data structure of the LUN to indicate that none of the LUN is yet initialized. As used herein, "initialized" refers to consistency of stored data and related redundancy information. As noted above, prior techniques

insisted on either initializing in the entire LUN to zero or at a minimum initializing the redundancy information portion of the LUN to values consistent with the otherwise un-initialized data portions. The boundary parameter value is used to distinguish portions of a LUN that have been initialized from those portions yet to be initialized. Element **208** is then operable to enable processing I/O requests in parallel with the initialization of the LUN about to commence. Processing of I/O requests in accordance with the present invention are discussed in further detail herein below.

Lastly, element **210** is operable to initialize the LUN as a background process operating in parallel with I/O request processing. As discussed further herein below, I/O request processing proceeds in accordance with the present invention in cooperation with simultaneous, background initialization of the LUN redundancy information. The boundary parameter noted above is used to aid the I/O request processing methods of the present invention in determining how to properly process a particular I/O request. Additional details of this cooperative processing are discussed herein below. This completes the method of the present invention for creation or definition of a new LUN.

LUN Initialization Method

FIG. **3** is a flowchart describing processing of element **210** in additional detail. Specifically, element **210** performs initialization of the redundancy information in a LUN to assure integrity and consistency of the redundancy information with the otherwise un-initialized data of the LUN. Those skilled in the art will recognize that such redundancy information initialization is dependent on the type of RAID management associated with the LUN. Specifically, RAID level 1 (mirroring) redundancy information initialization entails copying the un-initialized data to the corresponding mirrored position in the LUN, or in the alternative, initializing both the data and redundancy portion to zeros. By contrast, initialization of redundancy information in a RAID level 5 LUN entails reading the otherwise un-initialized data in each stripe of the LUN, computing the Exclusive OR parity of the un-initialized data, and storing such generated parity information in appropriate blocks of the stripe.

In particular, element **300** is first operable to set a local variable CURRENT to point to the first physical portion of storage space in the LUN (i.e., to the start of the LUN). Elements **302** through **310** are then iteratively operable to initialize each subsequent portion of the LUN. Those skilled in the art will recognized that each "portion" may constitute one or more stripes where appropriate to the RAID management or may constitute an arbitrary defined buffer size as appropriate for the particular RAID management level associated with the LUN.

Element **302** is specifically operable to initialize redundancy information for the portion of the LUNs presently pointed to by the local variable CURRENT (e.g., the current portion of the LUN). Element **304** is then operable to determine whether the initialization process performed by element **302** detected any failure in a disk drive of the LUN. As noted above, element **302** may read and/or write various blocks of the LUN for purposes of initializing associated redundancy information. If any of the read or write operations result in detection of failure of a disk drive, as detected by element **304**, processing continues with element **314** to mark the LUN as operable though only in degraded mode. This then completes processing of element **210**.

If no such drive failure is detected, element **306** is next operable to adjust the boundary parameter value in the LUN meta-data structure to point past the CURRENT portion of the LUN (i.e., to point to the next portion of LUN yet to be initialized). Element **307** then saves state information regarding the progress of the LUN initialization. This saved data (also referred to herein as checkpoint data) is used later if the initialization process is resumed after being interrupted (i.e., by a power loss in the storage system).

Element **308** is then operable to determine whether more portions of the LUN remain to be initialized. If so, element **310** is next operable to set the local variable CURRENT to the next portion of the LUN to be initialized. Processing then continues by looping back to element **302** until all portions of LUN had been so initialized or until a disk drive failure is detected by processing of element **304**. In the preferred embodiment, a delay may be inserted in the processing of the initialization to permit resources of the storage system to be utilized for processing host requests rather than being totally consumed by the initialization sequence. In the alternative, a periodic check can be made to determine if such host I/O requests are pending or queued. When such requests are found to be queued the initialization process may then pause briefly to permit processing of those requests. Such design choices of multitasking and processor time sharing are well known to those skilled in the art.

Element **312** is finally operable when all portions of the LUN have been initialized to mark the LUN has fully initialized allowing all I/O requests to be processed in accordance with normal I/O processing for LUN. Processing of I/O requests in cooperation with the initialization process of FIG. **3** is discussed in additional detail herein below.

The processing of FIG. **3** may also be entered in response to a resumption of the initialization process following interruption of a previous initialization. As noted above, the method of FIG. **3** saves checkpoint data as it progresses in the initialization process. This checkpoint data is used in element **316** to restore the previous state of the initialization process. In this manner, a previously interrupted initialization process may be resumed following such an interruption,

I/O Request Processing Method

FIG. **4** is a flowchart describing the operation a method of the present invention for processing I/O requests in parallel with initialization as discussed above. In accordance with the present invention, I/O requests are handled in one of three manners depending on the state of the LUN and be progress of initialization of the LUN. Specifically, element **400** is first operable to determine whether LUN is operating in degraded mode in response to sensing of failure of one or more disk drives. If the LUN is presently operating in such a degraded mode, element **410** is operable to perform the requested I/O operations in the degraded mode of the LUN. Such degraded mode operation of a disk array LUN is well known to those skilled in the art. In general, such degraded mode operation entails carefully updating data and/or associated redundancy information in such a manner as to ensure that no data will be lost so long as no further disk failures occur.

If the LUN is not operating in degraded mode but rather is operable in normal mode, element **402** is next operable to determine whether the LUN has been fully initialized as described above. If the initialization to the LUN has completed as described above with respect to FIG. **3**, element **406** is next operable to process the I/O request in a normal fashion in accordance with the RAID management level associated with the LUN.

If the LUN is not yet fully initialized, element **404** is operable to determine whether the particular I/O request requires access exclusively to data and redundancy information above the presently set boundary parameter value.

As noted above, the boundary parameter value in the LUN the meta-data structure is indicative of the progress of the initialization of the LUN as described above with respect to FIG. **3**. If this I/O request access data or redundancy information completely above the present boundary parameter value, element **406** is operable as described above to process the I/O request in a normal fashion in accordance with the RAID management level associated with the LUN.

If the data to be accessed is not fully above the boundary, element **405** is next operable to determine if the data to be accessed is fully below the boundary. If so, the entire request is in the un-initialized area of the LUN and element **408** is operable to perform the I/O request in a manner cognizant of the un-initialized information in the LUN. Specifically, the I/O request is processed by generating new redundancy information consistency with the data to be written by the I/O request (and with any related old, un-initialized data remaining in affected areas of the LUN).

If the data to be accessed is neither fully above nor fully below the present boundary but rather spans the boundary, element **407** is operable to requeue the I/O request for later processing. The I/O request is requeued since it is likely that the initialization processing will proceed to the point past the data to be accessed in the near future. This brief delay is found to be relatively unimportant in the overall processing by the storage system. Such instances occur rarely. In the alternative, those skilled in the art will recognize that an I/O request that spans the present boundary may also be processed as in element **408** above. Such processing would conservatively process the I/O request as though it were below the boundary in its entirety. Such equivalent design choices are well known to those skilled in the art.

It all four cases, that is processing I/O requests by element **406**, element **407**, element **408**, or element **410**, processing of the I/O request is completed.

Those skilled in the art will recognize that the processing of element **408** to generate new redundancy information is performed in accordance with the particular RAID management level associated with the LUN. For example, where the LUN is managed in accordance with RAID level 1 mirroring, generation of new redundancy information entails writing of the user requested data and duplicating the data in the associated mirror portion of the LUN. By contrast, where the LUN is managed in accordance with RAID level 5 techniques, processing of element **408** entails generation of new redundancy information by Exclusive OR parity operations including the data modified by the I/O write request and any associated old data remaining on the LUN in the related stripe. It all such cases, operation of element **408** leaves the redundancy information associated with the I/O write request in a consistent state such that it maybe relied upon in processing of subsequent read or write requests.

Method for Handling Disk Failure During Initialization

FIG. **5** is a flowchart describing a method of the present invention operable in response to sensing the restoration of operation of a previously failed disk drive. As noted above, initialization of the LUN in accordance with the present invention is terminated when the initialization sequence senses failure of one or more disk drives in the LUN. The LUN is marked so that I/O request processing may continue though in a degraded mode of operation. When the failed disk drive has been restored to normal operation (i.e., repaired and/or replaced), standard RAID management regeneration techniques are used to assure integrity and reliability of the data stored on LUN. Such regeneration also serves to complete the previously terminated initialization of the LUN by assuring that all data and associated redundancy information in the LUN are consistent.

In particular, element **500** is operable in accordance with the RAID management technique presently assigned to the LUN to regenerate all LUN data and redundancy information affected by replacement of the failed disk drive. Element **502** is then operable to mark the LUN as restored to normal operation as distinct from degraded mode operation. Lastly, element **504** is operable to mark the LUN as fully initialized because the regeneration techniques of the RAID management have accomplished the purpose of initialization: namely, to make all redundancy information consistent with the data stored in the LUN.

FIG. **6** is a block diagram depicting the significant functional elements operable within a controller **102** in accordance with the present invention. Those skilled in the art will recognize the functional decomposition of FIG. **6** as merely one exemplary decomposition of the functional elements. Many equivalent functional decompositions of the processing of the present invention will occur to those skilled in the art.

A LUN creator element **600** is tightly coupled to a LUN initialization element **602**. LUN creator **600** creates a newly defined LUN in accordance with input from a user of the system. The user defines a number of parameters for the LUN architecture including the RAID management level to be used with the LUN and physical, geometric mapping of data on the LUN.

I/O request processor **606** processes I/O requests received from attached host systems. As noted above, the methods and structure of the present invention permit LUN initializer **602** and I/O request processor **606** to carry on processing in parallel. LUN initializer **602** maintains status information in boundary value **604** indicative of progress in the initialization process. I/O request processor **606** inspects the boundary value **604** to determine whether the I/O request may be processed normally or must be processed in a special manner due to the parallel initialization processing.

I/O request processor **606** is therefore comprised of a variety of processing elements for processing of I/O requests. Normal processing **608** processes I/O request in the normal course of storage management associated with the LUN. Normal processing **608** is invoked when the I/O request accesses only data that resides fully within the area of the LUN for which initialization has completed. Where the data to be accessed by an I/O request resides fully outside the initialized portion of the LUN, full stripe processing **610** is invoked. Full stripe processing **610** assures integrity of the data to be written to the un-initialized area of the LUN by performing only full stripe writes. The full stripes include newly generated redundancy information generated by redundancy generator element **612** (i.e., XOR parity generation). Writing of a full stripe assures integrity of the data and associated redundancy information in the stripe despite the fact that the LUN initializer **602** has not yet processed through that area of the LUN. Lastly, as noted above, I/O request processor **606** may include a deferred processing element **614** to simply requeue an I/O request for later processing when the LUN initializer **602** has proceeded through the area affected by the I/O request. For example, an I/O request that spans an area of the LUN including the present boundary value **604** (i.e., crosses the present boundary of initialization processing), may be delayed for later processing by requeuing it. Later, when the request is again processed by I/O request processor **606**, the initialization of the LUN may have passed by the affected area permitting the I/o request to be processed as a normal request.

Drive failure/restoration detector element **616** monitors the status of the disk drives to determine whether a drive in

the LUN is has failed during the initialization processing. If so, LUN initializer **602** is terminated. When the failed disk drive is restored to normal operation, LUN regeneration element **618** is notified to commence processing to regenerate the data and redundancy information of the LUN in accordance with the RAID storage management level assigned to the LUN. This regeneration by standard RAID techniques serves the dual purpose of completing the initialization sequence that was terminated upon sensing of the drive failure.

While the invention has been illustrated and described in detail in the drawings and foregoing description, such illustration and description is to be considered as exemplary and not restrictive in character, it being understood that only the preferred embodiment and minor variants thereof have been shown and described and that all changes and modifications that come within the spirit of the invention are desired to be protected.

What is claimed is:

1. A method operable in a RAID storage system for enabling immediate access to said RAID storage system comprising the steps of:

creating a RAID level 5 logical unit within said RAID storage system;

initializing said logical unit in response to creation of said logical unit;

processing I/O requests directed to said logical unit in parallel with the step of initializing, wherein the step of initializing comprises steps of

maintaining a boundary parameter value indicative of the progress of the initialization of said logical unit wherein said boundary parameter value indicates a physical position in said logical unit above which initialization is complete and below which initialization is incomplete,

saving checkpoint data periodically during processing of the step of initializing indicative of progress of the step of initializing,

interrupting processing of the step of initializing, and

resuming the step of initializing using said checkpoint data, and the step of processing comprises steps of

determining whether an I/O request to be processed includes access to storage space of said logical unit below said boundary parameter value,

processing said I/O request normally in response to determining that said I/O request does not include access to storage space below said boundary parameter value,

processing said I/O request in a manner so as to assure integrity of redundancy information associated with data of said I/O request in response to determining that said I/O request requires access to storage space below said boundary parameter value,

requeuing said I/O request for later processing in response to a determination that the data to be accessed crosses said boundary parameter value.

2. The method of claim **1** further comprising the steps of:

sensing failure of a failed disk drive of said logical unit during the step of initialization; and

terminating the step of initialization in response to sensing of said failure.

3. The method of claim **2** further comprising the steps of:

sensing restoration of operation of said failed disk drive; and

regenerating information stored in said logical unit to complete initialization of said logical unit in response to sensing of said restoration of operation.

4. The method of claim **2** wherein the step of processing I/O requests includes the step of:

operating said logical unit in a degraded mode to continue processing of said I/O requests.

5. The method of claim **1** wherein the step of processing said I/O request to assure integrity includes the step of:

generating new redundancy information from data associated with said I/O request regardless of un-initialized redundancy information presently stored on said logical unit below said boundary parameter value.

6. A RAID storage system comprising:

a logical unit creator to define a RAID level 5 logical unit in said storage system;

a logical unit initializer to initialize said logical unit by assuring consistency of redundancy information and data in said logical unit, save checkpoint data periodically during initialization indicative of progress of the initialization, interrupt the initialization, and resume the initialization using said checkpoint data; and

an I/O request processor to process I/O requests directed to said logical unit from attached host systems, wherein said logical unit initializer is operable in parallel with said I/O request processor, said I/O request processor further configured for determining whether an I/O request to be processed includes access to storage space of said logical unit below said boundary parameter value, processing said I/O request normally in response to determining that said I/O request does not include access to storage space below said boundary parameter value, processing said I/O request in a manner so as to assure integrity of redundancy information associated with data of said I/O request in response to determining that said I/O request requires access to storage space below said boundary parameter value, requeuing said I/O request for later processing in response to a determination that the data to be accessed crosses said boundary parameter value.

7. (AMENDED) The system of claim **6** wherein said I/O request processor further comprises:

a second I/O request processor to process I/O requests by generating new redundancy information in response to determining that information required for processing said I/O request has not been initialized as indicated by said boundary parameter indicator wherein said second I/O processor includes:

a redundancy information generator for generating new redundancy information from data supplied in said I/O request regardless of un-initialized redundancy information presently stored is said logical unit.

8. The system of claim **6** wherein said logical unit initializer includes:

a disk drive failure detector to detect failure of a failed disk drive in said logical unit during operation of said logical unit initializer wherein said disk drive failure detector terminates operation of said logical unit initializer in response to detection of failure of said failed disk drive in said logical unit.

9. The system of claim **8** further comprising:

a disk drive replacement detector for detecting repair or replacement of said failed disk drive; and

a logical unit regenerator, coupled to said disk drive replacement detector, for regenerating data and redundancy information affected by failure of said failed disk drive.

10. A storage controller in a RAID storage system for enabling immediate access to said RAID storage system, said storage controller comprising:

means for creating a RAID level 5 logical unit within said RAID storage system;

means for initializing said logical unit in response to creation of said logical unit; and

means for processing I/O requests directed to said logical unit in parallel with operation of said means for initializing, wherein the means of initializing comprises

means for maintaining a boundary parameter value indicative of the progress of the initialization of said logical unit wherein said boundary parameter value indicates a physical position in said logical unit above which initialization is complete and below which initialization is incomplete,

means for saving checkpoint data periodically during processing of the initialization indicative of progress of the initialization,

means for interrupting processing of the means for initializing, and

means for resuming the initialization using said checkpoint data, and the means for processing comprises

means for determining whether an I/O request to be processed includes access to storage space of said logical unit below said boundary parameter value,

means for processing said I/O request normally in response to determining that said I/O request does not include access to storage space below said boundary parameter value,

means for processing said I/O request in a manner so as to assure integrity of redundancy information associated with data of said I/O request in response to determining that said I/O request requires access to storage space below said boundary parameter value,

means for requeuing said I/O request for later processing in response to a determination that the data to be accessed crosses said boundary parameter value.

11. The controller of claim **10** further comprising:

means for sensing failure of a failed disk drive of said logical during operation of said means for initializing; and

means for terminating the said means for initializing in response to sensing of said failure.

12. The controller of claim **11** further comprising:

means for sensing restoration of operation of said failed disk drive; and

means for regenerating information stored in said logical unit to complete initialization of said logical unit in response to sensing of said restoration of operation.

13. The controller of claim **11** wherein the means for processing I/O requests includes:

means for operating said logical unit in a degraded mode to continue processing of said I/O requests.

14. The controller of claim **10** wherein the means for processing said I/O request to assure integrity includes:

means for generating new redundancy information from data associated with said I/O request regardless of un-initialized redundancy information presently stored on said logical unit below said boundary parameter value.

* * * * *

US005974544A

United States Patent [19]

Jeffries et al.

[11] Patent Number: 5,974,544

[45] Date of Patent: Oct. 26, 1999

[54] METHOD AND CONTROLLER FOR DEFECT TRACKING IN A REDUNDANT ARRAY

[75] Inventors: **Kenneth Layton Jeffries**, Leander; **Craig S. Jones**, Austin, both of Tex.

[73] Assignee: **Dell USA, L.P.**, Round Rock, Tex.

[21] Appl. No.: **08/724,381**

[22] Filed: **Oct. 1, 1996**

Related U.S. Application Data

[63] Continuation of application No. 08/449,189, May 24, 1995, abandoned, which is a continuation of application No. 08/266,417, Jun. 27, 1994, abandoned, which is a continuation of application No. 07/808,330, Dec. 17, 1991, abandoned.

[51] **Int. Cl.**[6] **G06F 11/00**; G06F 12/12

[52] **U.S. Cl.** **713/1**; 710/10; 714/8

[58] **Field of Search** 395/651; 713/1; 710/10

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,924,331	5/1990	Robinson et al.	360/72.1
5,075,804	12/1991	Deyring	360/49
5,088,081	2/1992	Farr	369/54
5,111,444	5/1992	Fukushima et al.	369/58
5,146,571	9/1992	Logan	395/400
5,166,935	11/1992	Bish	371/21.6
5,166,936	11/1992	Ewert et al.	371/21.6
5,200,959	4/1993	Gross et al.	371/21.6
5,210,860	5/1993	Pfeffer et al.	395/183.18
5,220,569	6/1993	Hartness	371/37.7
5,233,618	8/1993	Glider et al.	395/182.04
5,237,553	8/1993	Fukushima et al.	369/58
5,249,279	9/1993	Schmenk et al.	395/825
5,271,018	12/1993	Chan	371/10.2
5,278,838	1/1994	Ng et al.	395/182.04
5,301,297	4/1994	Menon et al.	711/114
5,303,244	4/1994	Watson	395/182.03
5,313,626	5/1994	Jones et al.	395/182.03
5,390,187	2/1995	Stallmo	395/182.05
5,440,716	8/1995	Schultz et al.	711/114
5,502,836	3/1996	Hale et al.	395/182.01
5,519,844	5/1996	Stallmo	711/114
5,530,960	6/1996	Parks et al.	395/825
5,592,648	1/1997	Shultz et al.	711/114
5,619,723	4/1997	Jones et al.	395/823
5,708,769	1/1998	Stallmo	395/182.04

OTHER PUBLICATIONS

Que Publishing, Using Your Hard Disk, 1990, pp. 180–182.

Patterson et al, Introduction to Redundant Arrays of Inexpensive Disks (RAID), IEEE 1989, pp. 112–117.

Primary Examiner—Tariq R. Hafiz

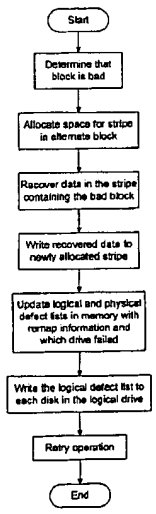
Assistant Examiner—John Q. Chavis

Attorney, Agent, or Firm—Skjerven, Morrill, MacPherson, Franklin & Friel, L.L.P.

[57] **ABSTRACT**

A disk controller for a disk drive array which maintains two representations of all drive defects. The controller maintains a logical defect list that is used to maintain the sector remapping structure when reconstructing redundancy information. The controller also maintains a physical defect list that is used to preserve known defect information on a physical disk basis. The physical defect list stores the defects even if the logical configuration of the disks changes. When the controller of the present invention determines that a block of data is bad, the controller allocates space for the respective stripe in an alternate block, recovers the data in the stripe and writes the recovered data to the newly allocated stripe. The controller then updates the remap tables in memory with the remap information. On each disk access, the controller searches the logical defect list to determine if the access involves one or more bad blocks. When a failed disk is replaced, the controller rebuilds the data from the failed drive using the remaining data and parity. The controller also uses both the logical and physical defect lists to unmap remapped sectors which were originally remapped due to defective sectors on the replaced disk drive.

19 Claims, 13 Drawing Sheets



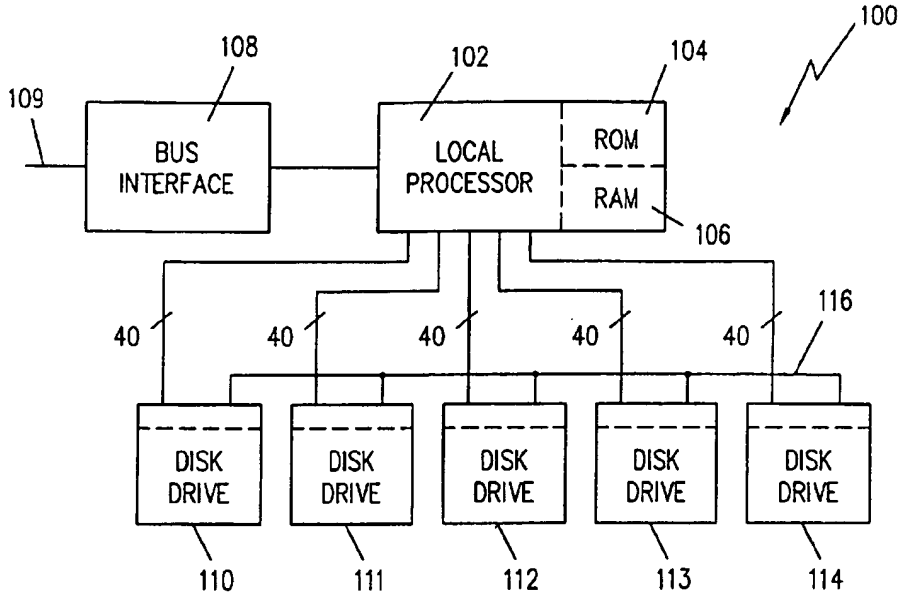

100
108
102
104
109
BUS INTERFACE
LOCAL PROCESSOR
ROM
RAM
106
40
40
40
40
40
116
DISK DRIVE
DISK DRIVE
DISK DRIVE
DISK DRIVE
DISK DRIVE
110
111
112
113
114

FIG. 1

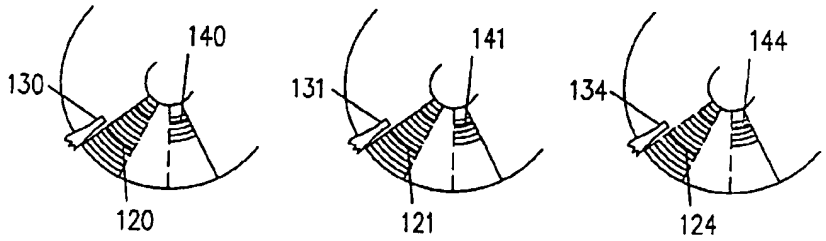

140
130
120
141
131
121
144
134
124

FIG. 2

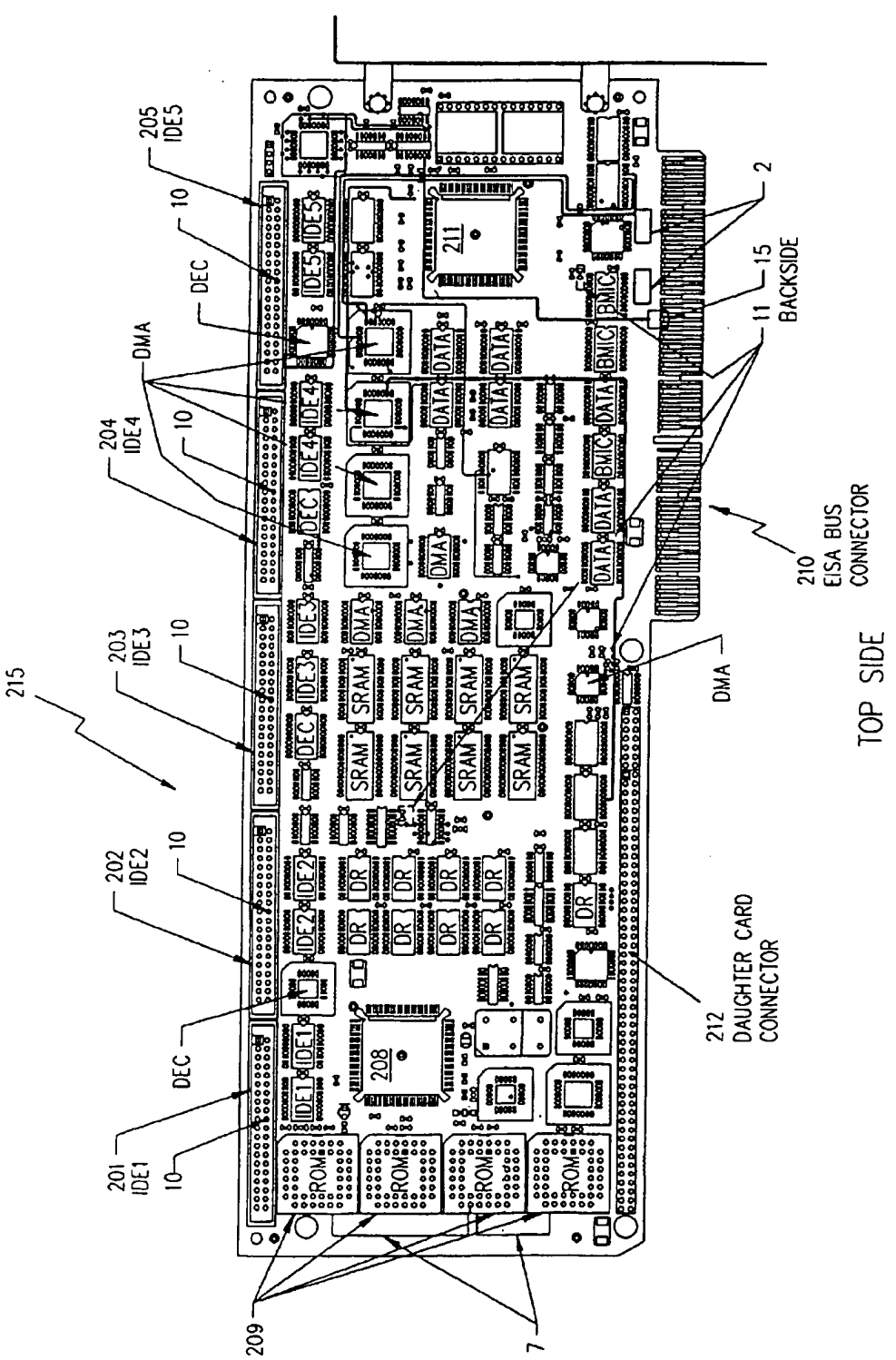

205
IDE5
10
DEC
DMA
204
IDE4
10
203
IDE3
10
215
202
IDE2
10
DEC
201
IDE1
10
209
7
211
208
2
15
BACKSIDE
11
210
EISA BUS
CONNECTOR
DMA
212
DAUGHTER CARD
CONNECTOR
TOP SIDE

FIG. 3

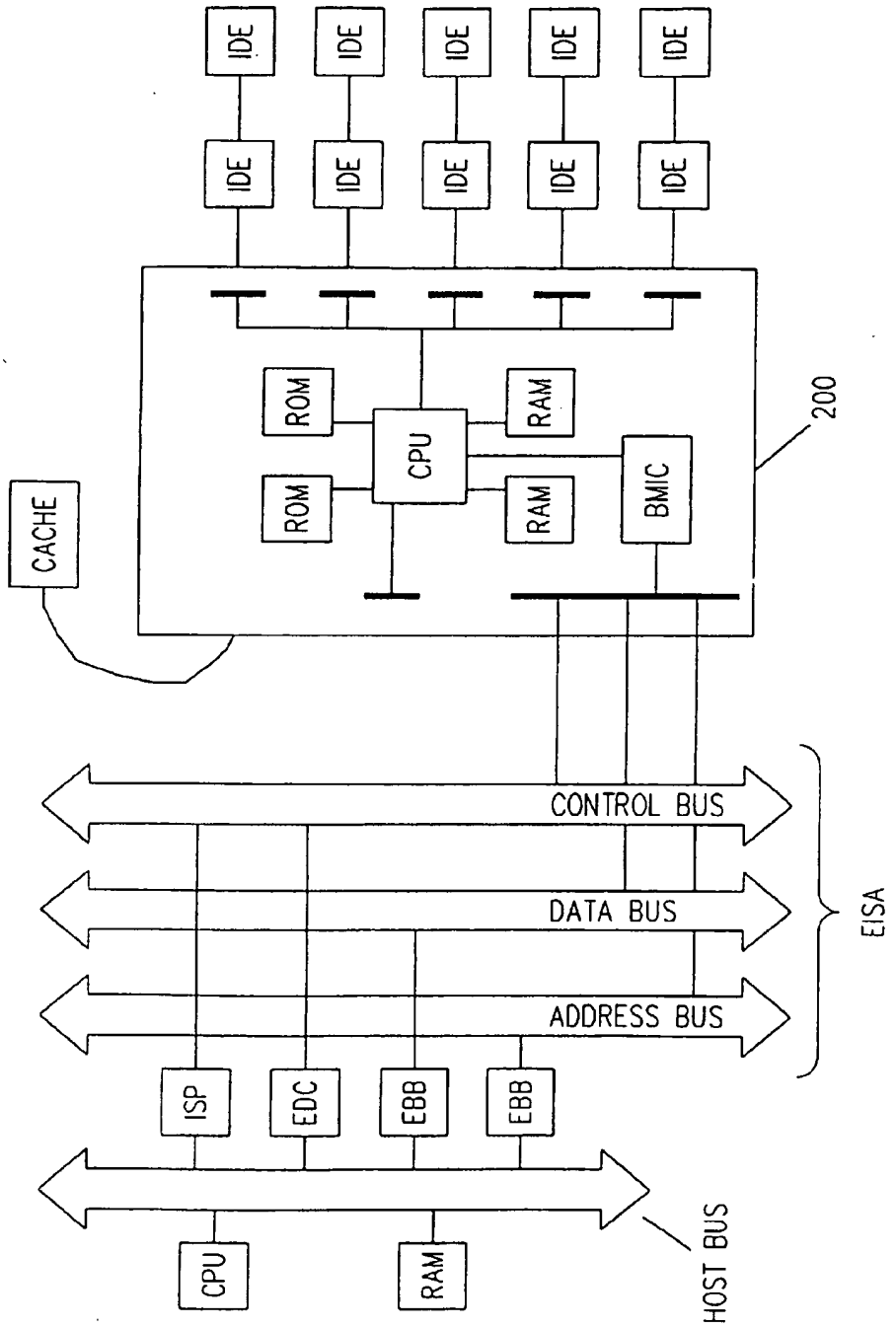

IDE
IDE
IDE
IDE
IDE
IDE
IDE
IDE
IDE
IDE
ROM
CPU
RAM
ROM
RAM
BMIC
CACHE
200
CONTROL BUS
DATA BUS
ADDRESS BUS
EISA
ISP
EDC
EBB
EBB
CPU
RAM
HOST BUS

FIG. 4

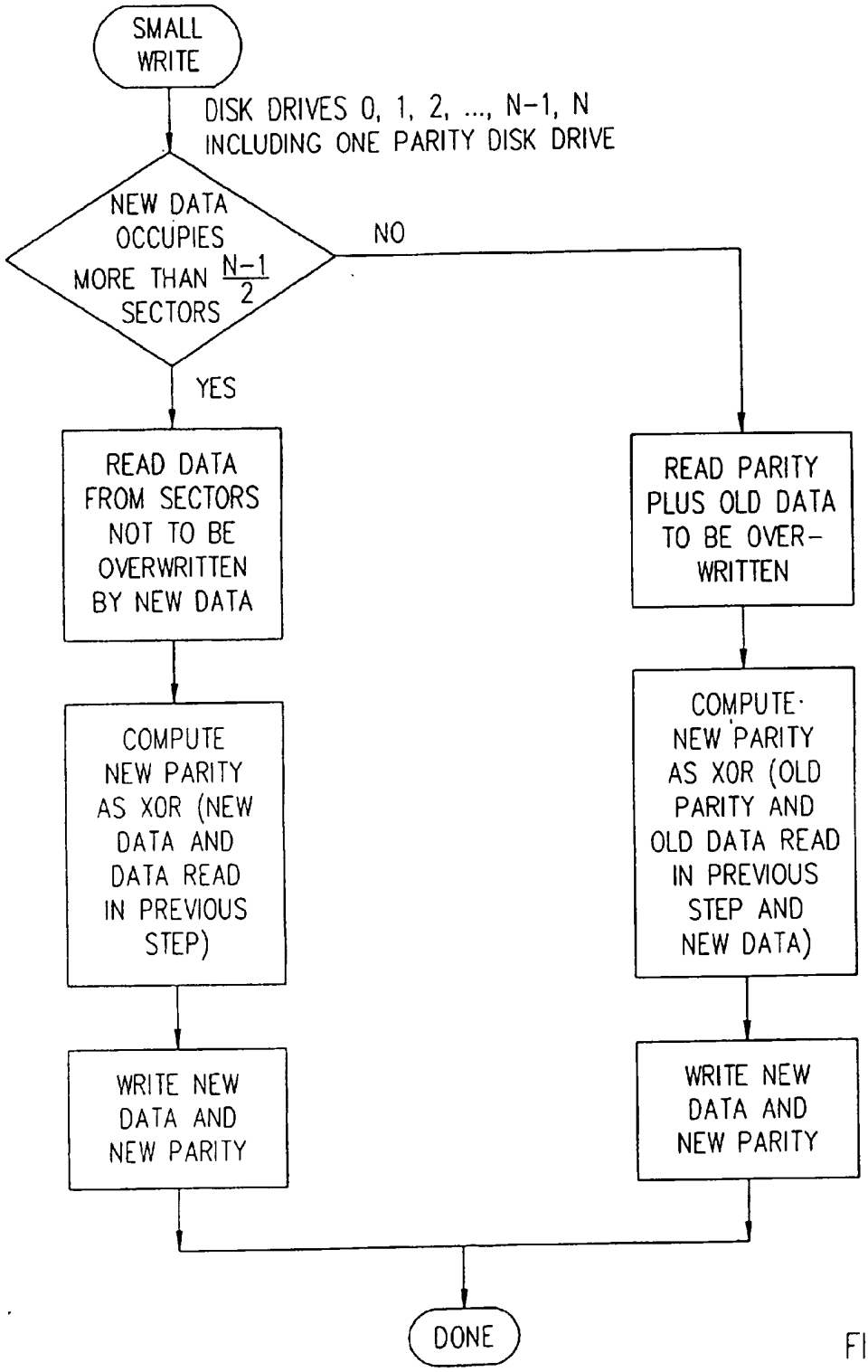

SMALL WRITE
DISK DRIVES 0, 1, 2, ..., N-1, N INCLUDING ONE PARITY DISK DRIVE
NEW DATA OCCUPIES MORE THAN $\frac{N-1}{2}$ SECTORS
NO
YES
READ DATA FROM SECTORS NOT TO BE OVERWRITTEN BY NEW DATA
READ PARITY PLUS OLD DATA TO BE OVER-WRITTEN
COMPUTE NEW PARITY AS XOR (NEW DATA AND DATA READ IN PREVIOUS STEP)
COMPUTE NEW PARITY AS XOR (OLD PARITY AND OLD DATA READ IN PREVIOUS STEP AND NEW DATA)
WRITE NEW DATA AND NEW PARITY
WRITE NEW DATA AND NEW PARITY
DONE

FIG. 5

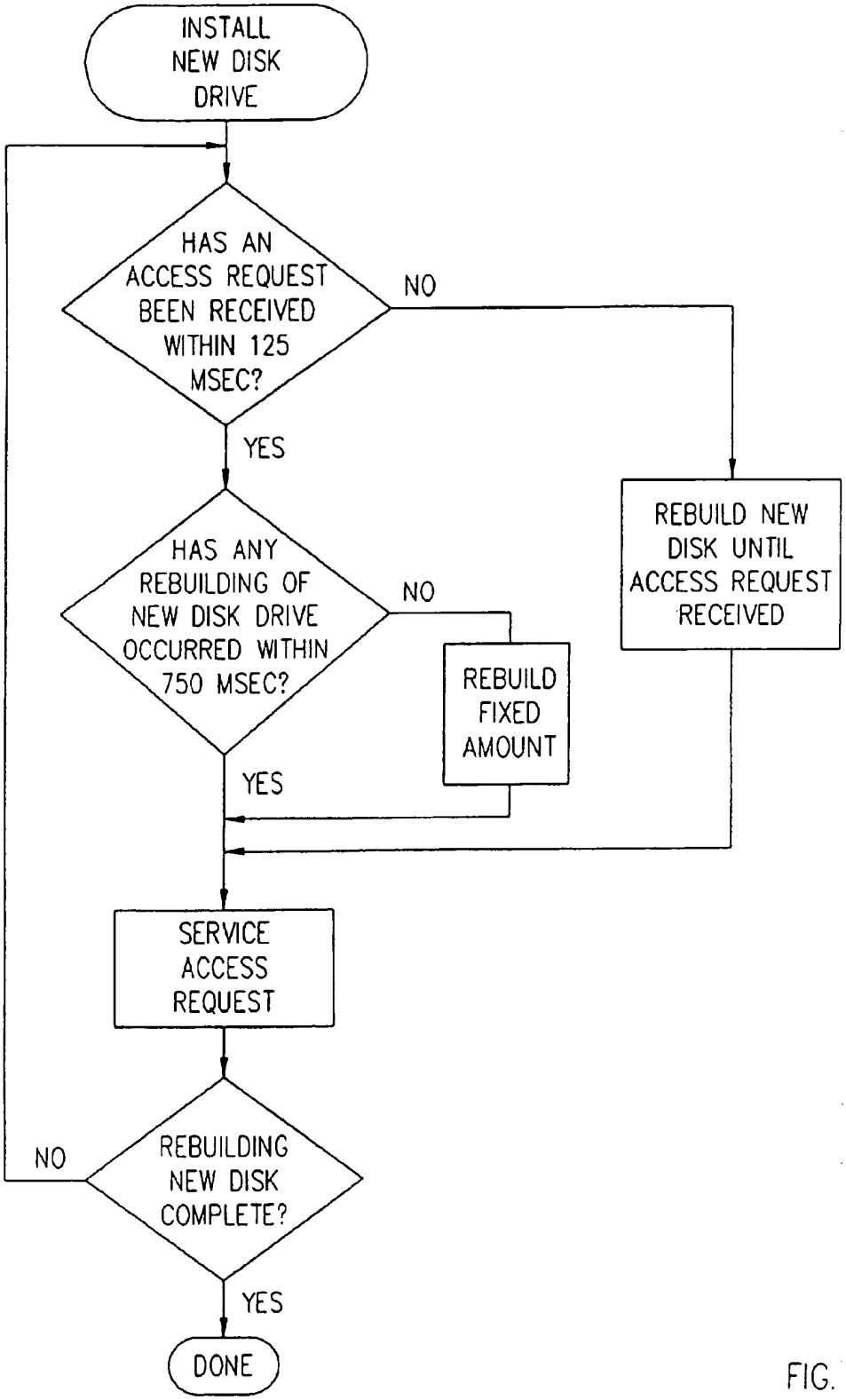

INSTALL NEW DISK DRIVE
HAS AN ACCESS REQUEST BEEN RECEIVED WITHIN 125 MSEC?
NO
YES
REBUILD NEW DISK UNTIL ACCESS REQUEST RECEIVED
HAS ANY REBUILDING OF NEW DISK DRIVE OCCURRED WITHIN 750 MSEC?
NO
REBUILD FIXED AMOUNT
YES
SERVICE ACCESS REQUEST
REBUILDING NEW DISK COMPLETE?
NO
YES
DONE

FIG. 6

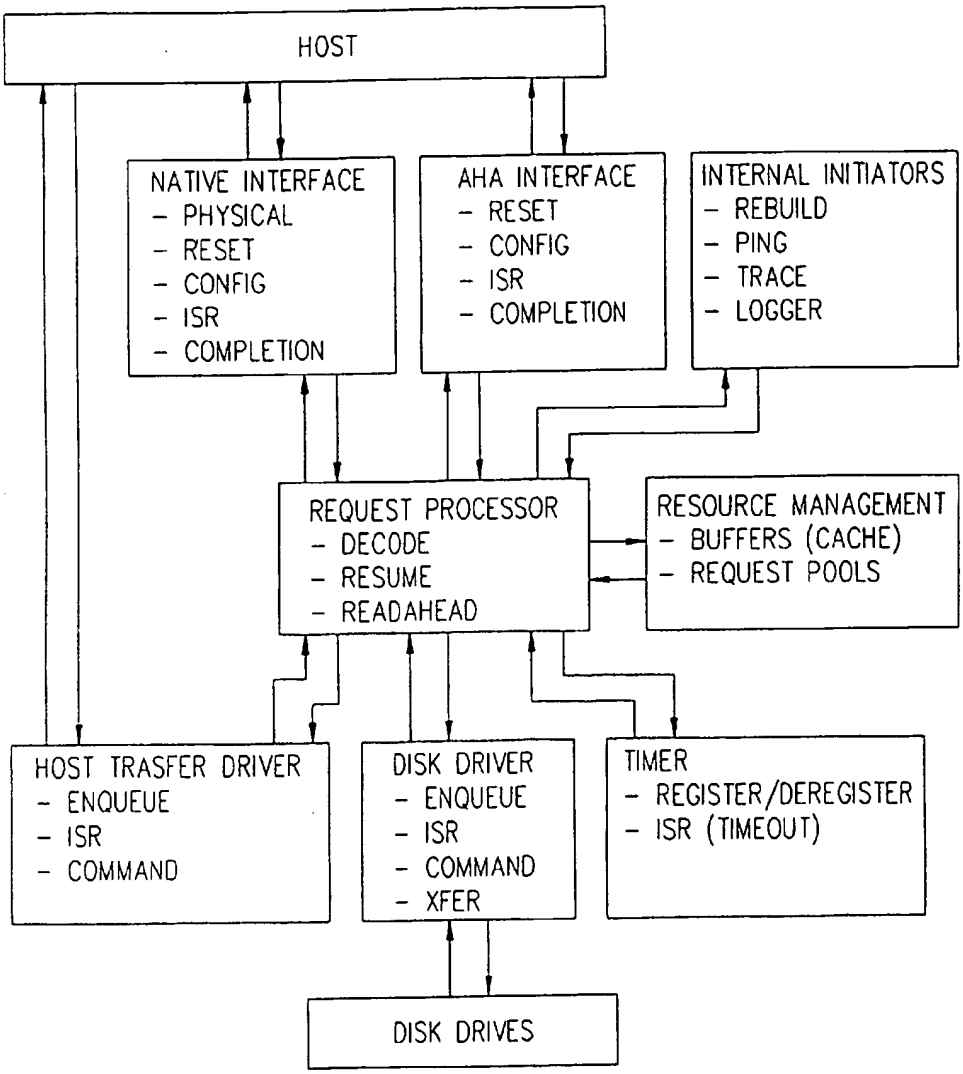

HOST
NATIVE INTERFACE
- PHYSICAL
- RESET
- CONFIG
- ISR
- COMPLETION
AHA INTERFACE
- RESET
- CONFIG
- ISR
- COMPLETION
INTERNAL INITIATORS
- REBUILD
- PING
- TRACE
- LOGGER
REQUEST PROCESSOR
- DECODE
- RESUME
- READAHEAD
RESOURCE MANAGEMENT
- BUFFERS (CACHE)
- REQUEST POOLS
HOST TRASFER DRIVER
- ENQUEUE
- ISR
- COMMAND
DISK DRIVER
- ENQUEUE
- ISR
- COMMAND
- XFER
TIMER
- REGISTER/DEREGISTER
- ISR (TIMEOUT)
DISK DRIVES

FIG. 7

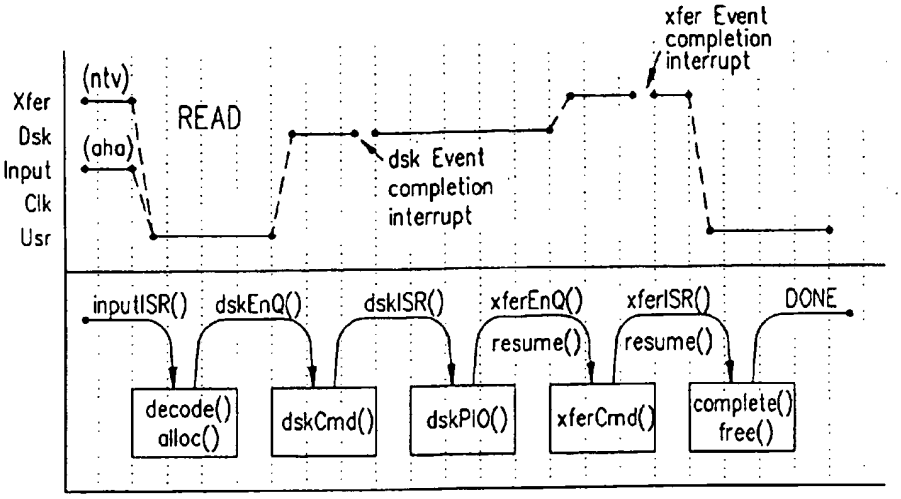


FIG. 8

SOFTWARE PRIVILEGE LEVEL

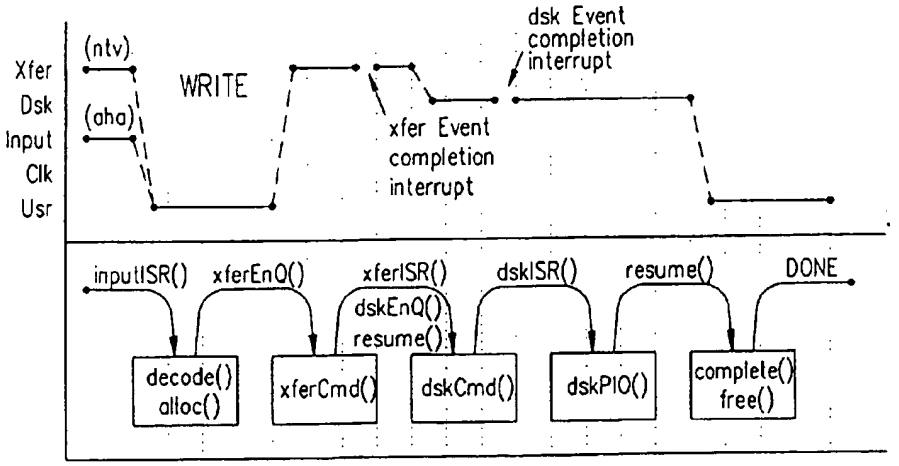


FIG. 9

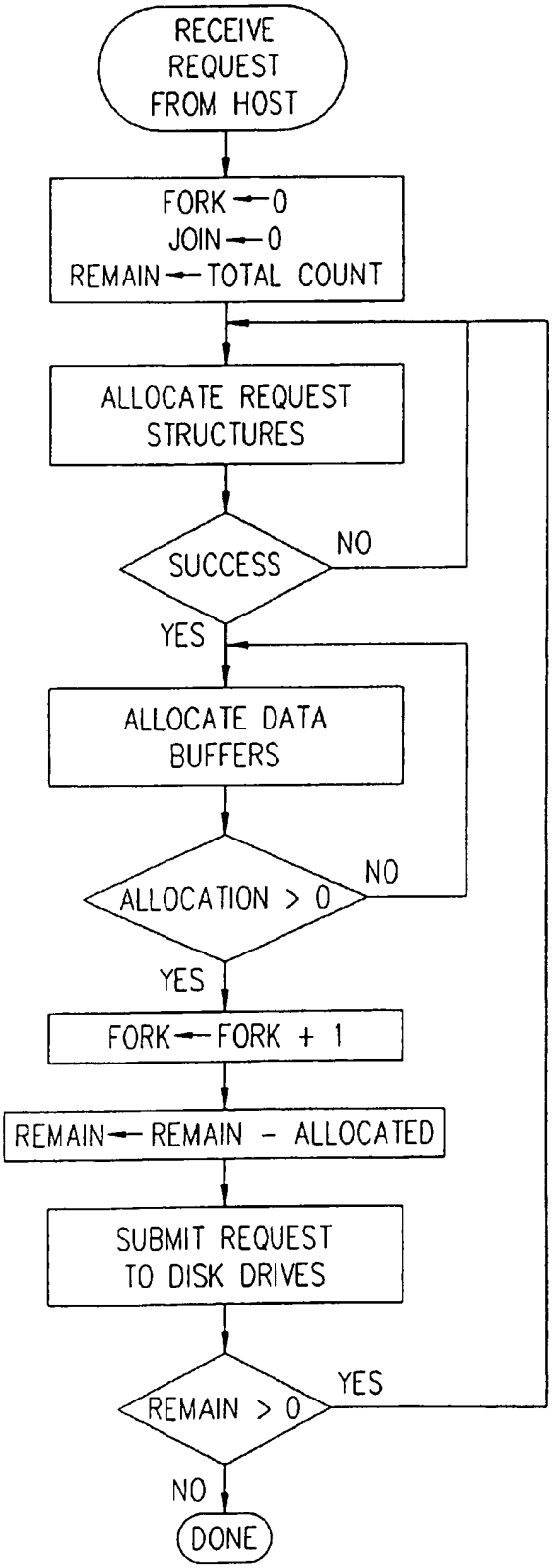

RECEIVE REQUEST FROM HOST
FORK ← 0
JOIN ← 0
REMAIN ← TOTAL COUNT
ALLOCATE REQUEST STRUCTURES
SUCCESS
NO
YES
ALLOCATE DATA BUFFERS
ALLOCATION > 0
NO
YES
FORK ← FORK + 1
REMAIN ← REMAIN - ALLOCATED
SUBMIT REQUEST TO DISK DRIVES
REMAIN > 0
YES
NO
DONE

FIG. 10

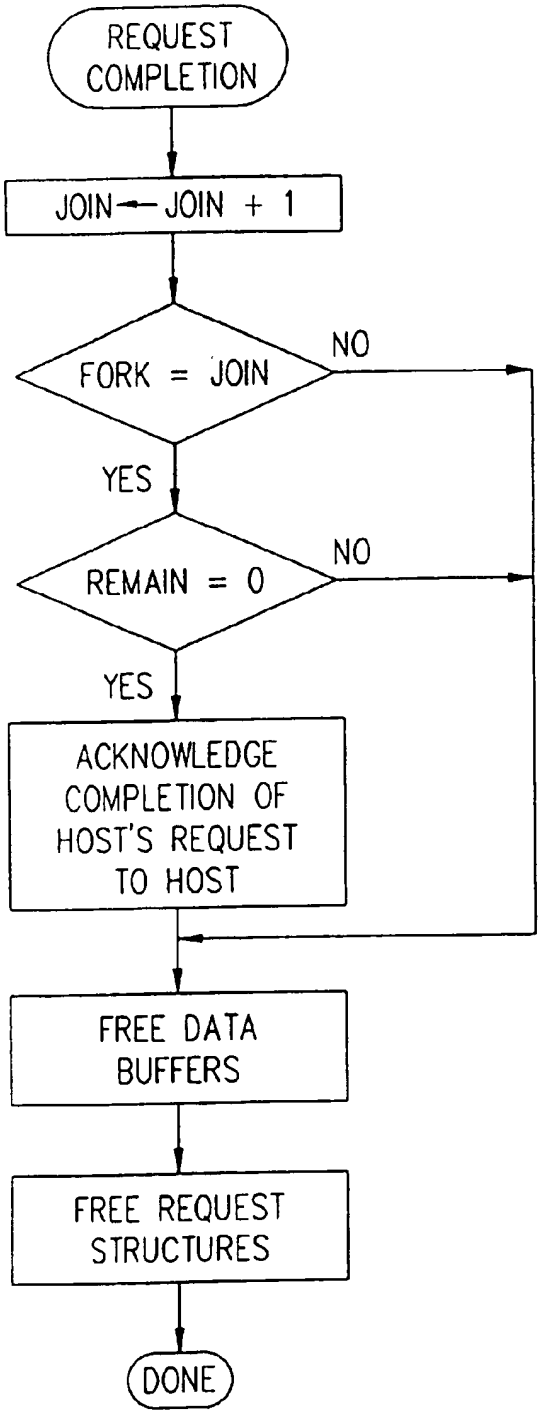

REQUEST COMPLETION
JOIN ← JOIN + 1
FORK = JOIN
NO
YES
REMAIN = 0
NO
YES
ACKNOWLEDGE COMPLETION OF HOST'S REQUEST TO HOST
FREE DATA BUFFERS
FREE REQUEST STRUCTURES
DONE

FIG. 11

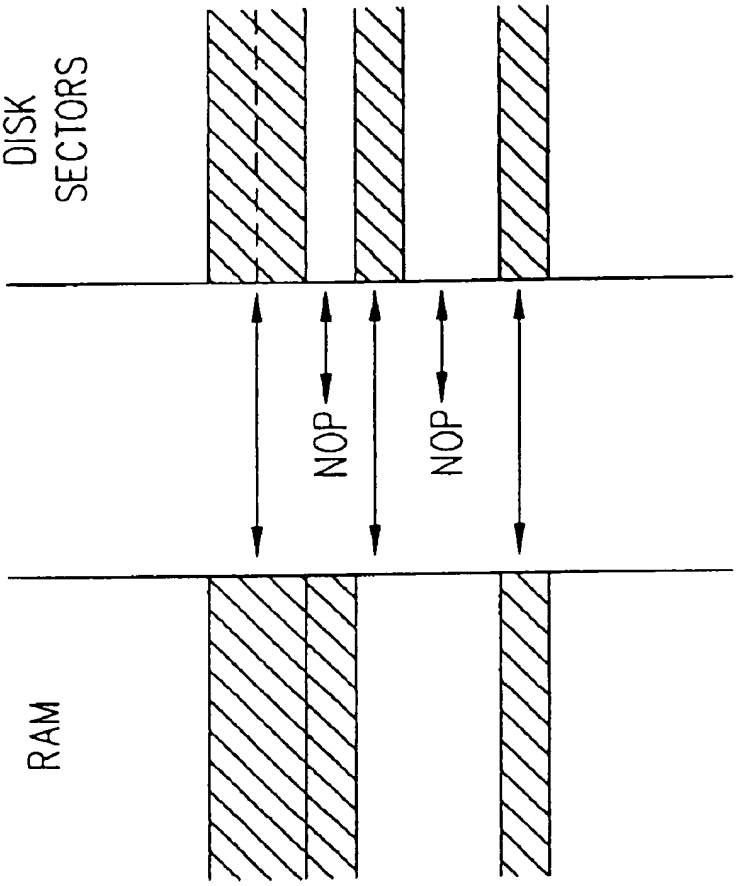
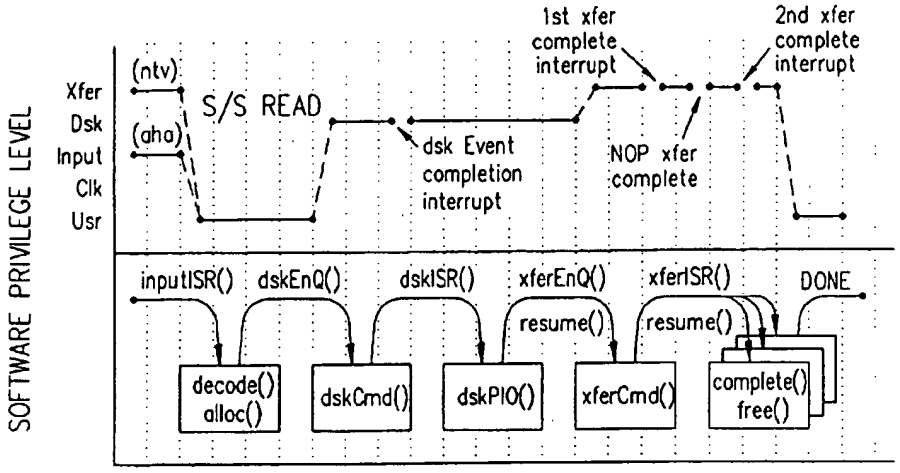


FIG. 12



Host Request

command	: S/S READ
host addr	: 0x1000
1st sector	: 100
count	: 3
drive	: 0

dsk Request

command	: READ
host addr	: 0x1001000
1st sector	: 100
count	: 6
drive	: 0

xfer Requests

command	: READ
host addr.	: 0x2000
DON addr	: 0x1001000
count	: 2
command	: NOP
host addr.	: –
DON addr	: 0x1001400
count	: 2
command	: READ
host addr.	: 0x3000
DON addr.	: 0x1001800
count	: 2

Completion called 3 times

0x1000 →

address	cnt
0x2000	2
-1	2
0x3000	2

DDA allocates:
- 6 Read Buffers
- 1 dsk Request
- 3 xfer Requests

fork = 3, join = 0

FIG. 13

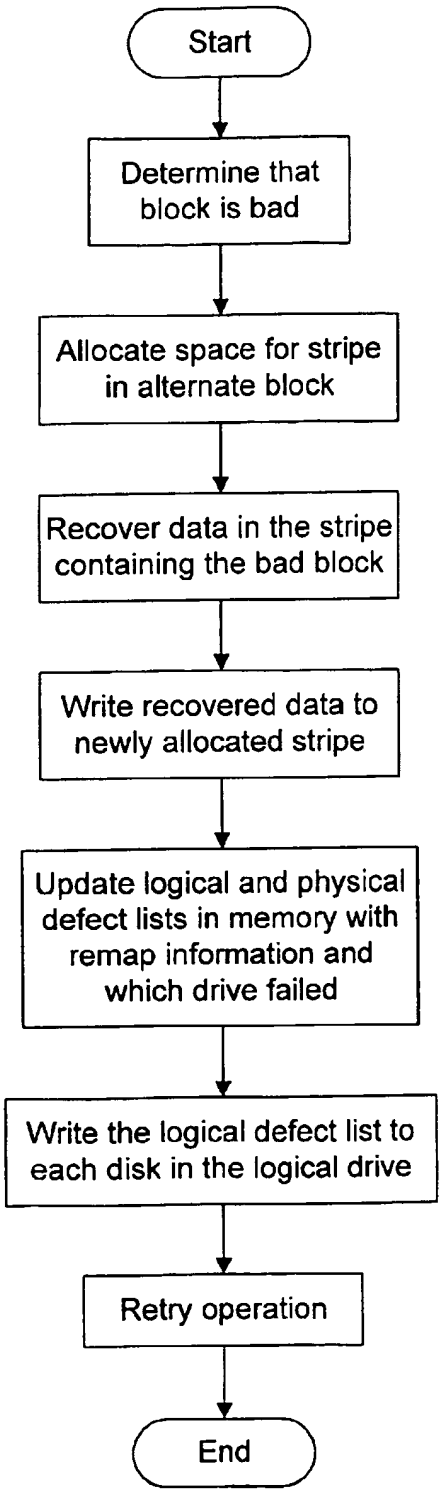

Start
Determine that block is bad
Allocate space for stripe in alternate block
Recover data in the stripe containing the bad block
Write recovered data to newly allocated stripe
Update logical and physical defect lists in memory with remap information and which drive failed
Write the logical defect list to each disk in the logical drive
Retry operation
End

Figure 14

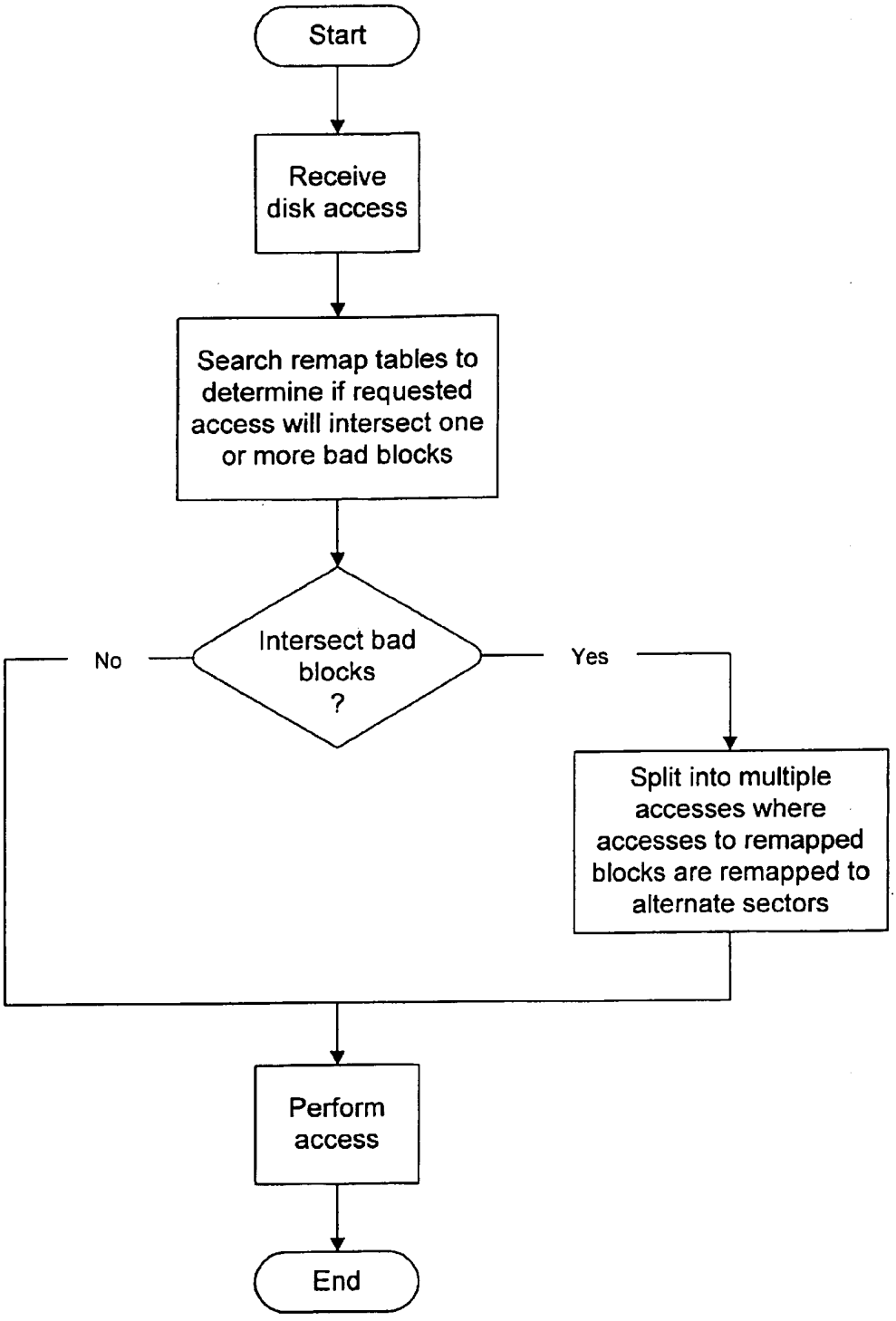

Start
Receive disk access
Search remap tables to determine if requested access will intersect one or more bad blocks
Intersect bad blocks ?
No
Yes
Split into multiple accesses where accesses to remapped blocks are remapped to alternate sectors
Perform access
End

Figure 15

METHOD AND CONTROLLER FOR DEFECT TRACKING IN A REDUNDANT ARRAY

This application is a continuation of application Ser. No. 08/449,189, filed May 24, 1995, now abandoned, which is a continuation of application Ser. No. 08/266,417, filed Jun. 27, 1994, abandoned, which is a continuation of application Ser. No. 07/808,330 filed on Dec. 17, 1991, abandoned.

CROSS-REFERENCE TO OTHER APPLICATIONS

The following applications of common assignee contain at least some drawings in common with the present application, and are believed to have effective filing dates identical with that of the present application, and are all hereby incorporated by reference:

- Ser. No. 07/809,452, filed Dec. 17, 1991 entitled "Disk Controller with Dynamic Sector Remapping" (DC-183);
- Ser. No. 07/810,278, filed entitled "Disk Drive Array with Dynamic Read Minimization" (DC-187);
- Ser. No. 07/810,790, filed Dec. 17, 1991, entitled "Disk Drive Array with Request Fragmentation" (DC-188);
- Ser. No. 07/808,841, filed Dec. 17, 1991, entitled "Disk Drive Array with Efficient Background Rebuilding" (DC-189);
- Ser. No. 07/808,716, filed Dec. 17, 1991, entitled "Disk Controller with Scatter/Scatter Transfer Operations" (DC-192);
- Ser. No. 07/808,801, filed Dec. 17, 1991, entitled "Disk Drive Array with Physical Mode" (DC-233);
- Ser. No. 07/810,277, filed Dec. 17, 1991, entitled "Disk Controller with Improved Read-Ahead Strategy" (DC-270); and
- Ser. No. 07/810,276, filed Dec. 17, 1991, entitled "Disk Drive Array with Periodic Activation of Physical Drives" (DC-272); all of which are hereby incorporated by reference. In particular, application DC-183 includes a very large appendix of source code, which provides even further detail on the contemplated best mode, and this appendix is expressly incorporated herein by reference.

BACKGROUND AND SUMMARY OF THE INVENTIONS

The present invention relates to electronic devices, and, more particularly, to disk drive memories for computers.

Mass Storage

In defining computer architectures, one of the basic trade-offs is memory speed versus cost: faster memory technologies cost more. Thus computer architectures usually include at least two levels of memory: a slow but large secondary memory, and a faster but smaller primary memory. The primary memory usually consists of volatile semiconductor random access memory (typically dynamic RAM or DRAM), and the secondary memory is usually nonvolatile (usually magnetic disk drives, hard and floppy, sometimes combined with optical disk drives, magnetic tape, etc.) Semiconductor RAM has access times on the order of 10–200 nanoseconds, with more expensive static RAM chips (SRAMs) faster than cheaper dynamic RAM chips (DRAMs); in contrast, magnetic hard disk drives (like other secondary memories) have much longer access times, due to the need for mechanical movement (the read/write head radial movement and the platter spin relative to the read/write, head).

Disk Drives

In a typical small hard disk drive, with a platter spinning at 3600 rpm, the latency time for a particular spot on the spinning platter to reach the read/write head after the head is positioned over the track containing the spot will vary between 0 and 16.7 milliseconds; and the seek time for the head to move to the correct track may vary from 1 to 30 milliseconds for small disk drives. Consequently, accessing a particular byte stored in a hard disk drive will take on the order of 20 milliseconds which is roughly five to six orders of magnitude greater than access time for semiconductor RAM. However, contiguous bytes in a track may be accessed in much less average time, because the head need not move (spin latency is 0); thus a disk drive may obtain read/write rates in the order of 10 million bits per second (e.g., 1 MBps) for contiguous bytes. Hence, for efficiency the central processor in a computer does not directly access secondary memory; but rather information in chunks is transferred between secondary and primary memory, and the central processor only deals with primary memory. In fact, in a virtual memory system pages of information are transferred between primary and secondary memory as the need arises to keep all physical addresses accessed in primary memory. This requires high speed transfer of large chunks of information between RAM and hard disk drives.

Risk of Failure

Hard disk drives (and all other mechanically involved secondary memory) have a problem of failure: both local platter defects and complete mechanical breakdown of the drive. Even with a mean time between failures of 100,000 or more hours for a disk drive, the failure rate and consequent loss of information may be too high for many computer users. Hence, various disk systems have been proposed to prevent loss of information upon disk drive failure.

Disk Arrays ("RAID")

In the 1980s a new technology was proposed to provide large disk capacity with very high reliability. This technology was originally referred to as a "redundant array of inexpensive disk drives," or "RAID". (The reference to "inexpensive" drives relates to the large high-data-rate drives commonly used in mainframe computers. These drives are much more expensive per byte than the small self-enclosed drives used in personal computers.) This technology is now also referred to generally as "disk drive array" technology.

This approach employs multiple physically separate disk drives, all accessed through a single array controller which makes the separate drives appear as a single composite drive. Typically the separate disk drives will be connected through a synchronization cable, so that the spinning platters of all drives are locked into the same rotational phase relative to each other. In different configurations, this approach can be used to achieve higher reliability, faster maximum data rate, and/or shorter maximum access time.

To achieve higher reliability, the information is stored so that each disk drive holds a check or parity byte for the corresponding bytes in the other disk drives. Thus if one disk drive fails, its information can be reconstructed from the information and parity on the other disk drives.

Information can also be stored in a stripe across the disk drives of a RAID system; this allows parallel read/writes by the disk drives and thereby increases information transfer speed. NCR's booklet "What are Disk Arrays?" (NCR 1990) illustrates various RAID type configurations.[1]

A further option is to store information redundantly in disk sectors at different angular positions; thus, for example, if a sector is repeated at four different positions on four different disks, the maximum rotational latency can be cut to 90 degrees equivalent instead of 360 degrees equivalent.

Queuing Access Requests

It may happen that a CPU central processing unit generates so many disk drive access requests within a small time interval that the requests cannot all be serviced immediately. (This is particularly likely when the CPU is operating in a multiprogramming environment.) In this case the disk drive controller (or CPU) queues these access requests and services them sequentially.

Sophisticated controllers for expensive disk drives may reorder the requests in queue (disk scheduling) to improve access efficiency; for example, the shortest-seek-time-first method would first service the request in the queue which involves the smallest distance of head movement, whereas the circular scan with rotational optimization method basically moves the head across the platter from outside track to inside track servicing requests in order of smallest head movement except for cases where radially-further-away sectors serviced first would lessen rotational latency. In these disk-scheduling controllers the request queue may be kept in an elevator queue (the requests are ordered as a function of sector radial distance) with each request identified by its associated handle. However, for inexpensive disk drives witn an Integrated Drive Electronics IDE (AT attachment design or ATA) or SCSI (small computer system interface) interface, as would be used in personal computers, the disk drive includes a controller on its circuit board to take care of hardware details such as motor control for head movement but would not include disk scheduling. An IDE disk drive only communicates with the CPU of a personal computer at a logic level rather than at a device level; this limits the CPU from disk scheduling because the IDE interface may include a mapping of the physical disk drive to appear as a different disk drive. For example, the use of 17 sectors per track was common for disk drives installed in IBM PCs in the early 1980s, but some more recent disk drives have used 40 sectors per track or even varying 35 to 49 sectors per track from spindle edge to outside edge; and these higher density disk drives can logically appear to have 17 sectors with multiple read/write heads. Consequently, IDE type disk drives have a problem of inefficient access, and a RAID with IDE disk drives compounds this problem.

Innovative Disk Array System

The present application discloses an innovative disk drive array system, including an innovative controller, which includes a large number of innovations. The following description summarizes some of the notable features of this system, including not only the claimed invention, but also many other features which are innovative or merely distinctive.

Dynamic Sector Remapping

Many disk/controller subsystems reserve storage to remap defective disk sectors. This remapping is statically or semi-statically defined at initial disk formatting. The system of the presently preferred embodiment (frequently referred to herein as the Dell Drive Array, or "DDA") automatically creates its initial remapping information and has the additional ability to dynamically remap "grown" defects.

Fork/Join/Remain for Request Synchronization

The system of the presently preferred embodiment uses a novel twist on a standard multi-processing software technique to implement multi-thread synchronization without the use of software critical regions or atomic read-modify-write cycles and without prior knowledge of the number of threads being created.

Dual-Queue/Semaphore Global Resource Protection

The system of the presently preferred embodiment uses a dual-queue mechanism to eliminate most software critical regions in the cache manager. This allows better system throughput by allowing the controller to perform the bulk of its lengthy data processing requirements using otherwise idle processor time while minimizing interrupt latency.

Dual Defect Lists

The system of the presently preferred embodiment maintains two independent representations of all drive defects. The first defect list is used to maintain the sector remapping structure when reconstructing redundancy information. This list is called the logical defect list and is stored in the remap data structure. The second list, called the physical defect list, is used to preserve known defect information on a physical disk basis. It allows defects to be maintained across logical configurations and is stored in a special reserved area which is always known even without a logical drive configuration.

Dynamic RMW Read Minimization

When reading data from a redundant array, there are always at least two ways to read the same data. This can be used in two ways: to improve performance and to handle errors. For mirrored arrays, large read requests can be accelerated. For guarded arrays, large read-modify-write (RMW) cycles can be accelerated.

The system of the presently preferred embodiment implements RMW acceleration by dynamically choosing read strategies which result in the smallest amount of data being transferred. The system of the presently preferred embodiment also uses redundancy to implement dynamic sector remapping transparent to host software and to provide traditional disk-failure recovery.

Request Decomposition to Allow Error Handling

The internal disk request queue provides a mechanism for associating multiple operations with a single disk request. This mechanism is used for two purposes: to attach internal operations unrelated to actual disk I/O, and to decompose complex requests to simple, fully restartable request sequences for error handling purposes.

Fragmentation to Atomic Operations

For queue management, all operation requests are fragmented down to "atomic," i.e. one-cycle, operations. This is highly advantageous for error handling: we never have to figure which phase of an operation we were in, because every operation is single-phase.

In order to ensure that the atomic relations stay in sequence, and that a sequence stays unbroken if it needs to, The system of the presently preferred embodiment uses "fence" markers in the queue, to fence a block of tasks. These markers are used as limits on the permissible queue-management operations, with rules which ensure that related sequences of atomic operations are kept together. Indeed, the small write discussed in connection with FIG. 5 includes both read and write operations, and permitting an intervening write to the data not to be written by the small write may change the parity and render the parity computed and written inaccurate. Thus fence markers can be used to keep out possibly disruptive intervening writes.

Fragmentation is an iterative list-manipulation process, which is repeated until the whole list is atomic.

In order to recover from an error condition, a computer system must be able to ascertain what went wrong. For example, if a single physical disk drive fails, it must be reset. The fragmentation of queue elements into atomic requests permits accurate recovery from error conditions without resorting to nested error handling routines.

Delay Strategy for Improved Responsiveness of Background Restores

The system of the presently preferred embodiment also implements a special strategy, for performing background data reconstruction, which attempts to minimize disk thrashing and maximize data bandwidth to the user while insuring completion of the reconstruction process.

To obtain high performance in a robust system, it is highly desirable to be able to perform rebuild operations in background. When soft error detection and correction is performed transparently to the user, the effective reliability of the system is increased.

However, a problem in achieving this is that there is likely to be a large seek time in moving the heads over to the rebuild data area (from the area accessed by the system). Thus, if the system begins a rebuild operation, a likely delay is superimposed on the delay time for the next access request from the host. If this conflict is badly managed, the disk's performance can be badly degraded by the large fraction of its time spent unavailable in transit. To avoid this, a disk system should avoid thrashing in and out of rebuild.

The system of the presently preferred embodiment provides two tunable parameters:

1) Amount of idle time before we initiate a rebuild;

2) maximum number of requests which may be processed before a rebuild operation is forced to occur.

A significant advantage of this innovative teaching is that the rebuild is guaranteed to complete within some determined time.

Note that this innovative teaching is not applicable only to drive arrays, but is also applicable (less advantageously) to a simple mirrored disk system.

Emulation of a Software Interface (INT13) in Hardware

In computer systems, it is common to see hardware interfaces simulated in software; however, the system of the presently preferred embodiment also emulates a software interface through its hardware interface. In the presently preferred embodiment, the input to the software interface (BIOS INT 13) is passed intact to the drive array controller, where portions of the BIOS characteristics are emulated by the DDA interface itself. (While this emulation is performed in the controller board's firmware, from the host's perspective it is emulated by the interface.)

Self-Assignment of Handles

Handles are devices that are frequently used across software and hardware interfaces to allow both sides of the interface to refer to common objects. Generally, these handles are created during the process of defining the object. As a result, two operations are usually required in such an interface: (1) creation/definition of the object, and (2) return of the handle to allow later references to the object. The system of the presently preferred embodiment eliminates the second step of this process by requiring the handle to be chosen by the side defining the object. This is an advantage for DDA since it allows the creation of a multi-request interface that imposes no performance penalty on host software (which only uses a single-request interface, but demands maximum performance).

Scatter/Scatter Read/Write Requests

"Scatter" and "gather" refer generally to common techniques in computer architecture: "scatter" is the transfer of a block of information from a contiguous set of locations to a discontiguous set of locations. Gather is the opposite process, ie. collecting information from a discontiguous set of locations for transfer to a contiguous set of locations.

Scatter and gather operations often arise in connection with a transfer of data across a boundary, e.g. from main memory to a peripheral. DMA controllers have included the capability for gather operations, to transfer a block of data from scattered locations of main memory out over the bus, or vice versa.

It is also suspected, although not known with certainty, that a Conner IDE drive currently in development allows a single I/O request to access discontinuous regions of the disk. It is also suspected, although not known with certainty, that some IBM drives may have included such a scatter capability. It is also not known whether these developments, if in fact they did occur, are prior art to any of the inventions in the present application.

Virtual memory operating systems, such as the UNIX Operating System, commonly result in a significant amount of memory scatter. In addition, the UNIX Operating System typically does not store contiguously on disk.

The system of the presently preferred embodiment provides "scatter/scatter" accesses, in which both the physical locations of data in host memory and the physical locations of data on the disks can be discontiguous. That is, the host can send a single request to launch such a scatter/scatter transfer. Arguments to such a transfer request would include: a pointer to a list of transfer counts and addresses in host memory containing the data to be transferred; the length of that list; and the starting logical address on the disk for transfer.

Note that the host need not know the configuration that the data array will have on the disk.

Skipped blocks in a scatter-scatter request are specified by a data address value of −1. Thus, when a block must be skipped, the controller enqueues a "nop" (no-operation)

request. Note that the presently preferred embodiment enqueues these nop requests, if needed, even if the data transferred is in contiguous addresses on the host memory side.

Any disk operation, in the presently preferred embodiment, is limited to a set maximum number of blocks of logical disk address space (currently 256). Thus, no scatter/scatter request can cover more than 254 skipped blocks.

The scatter-handling operations just described are implemented, in the presently preferred embodiment, using the controller's native mode described below.

As noted, virtual memory operating systems commonly result in a significant amount of memory scatter. In addition, the UNIX Operating System typically does not store contiguously on disk. Thus, this innovation is especially useful in UNIX systems.

Firmware Patching

The controller board of the presently preferred embodiment has only 256K of RAM, but has four specialized 128K EPROMs which are hard to change out. In this environment it is not easy to provide firmware flexibility.

In the controller board of the presently preferred embodiment, the firmware is made modular by heavy use of indirect calls. The firmware includes an INIT code section, called at board-reset time, which goes to a defined disk area (the "patch area") to pull up an updated set of address pointers. Changes to these pointers can be used to allow for configuration changes as well as for changes in code functionality.

Read-Ahead to Accommodate Multi-Thread Host Processor

The random-access speed of a disk drive is typically much slower than its serial-access speed, and very much slower than the clock speed of any associated CPU. Therefore, any lookahead strategy which succeeds in prefetching the data for any significant fraction of CPU access requests has the potential to improve performance. Disk drives are typically idle for a high percentage of the time, and this idle time can be used to perform lookahead operations.

However, lookahead reads are not necessarily advantageous: unless a sequential read operation is underway, a lookahead read would simply waste time.

One readahead strategy is simply to read a fixed amount ahead of the last data accessed. (In single disks, this was implemented simply by reading one or more tracks ahead.) This is a "dumb" strategy.

Another readahead strategy is to read ahead an amount dependent on the current read history. (This is an old strategy from mainframe databases.)

The system of the presently preferred embodiment uses a different readahead strategy: the controller keeps track of the last n reads (where n is a programmable parameter). If a new read comes in adjacent to any of the last n, a lookahead read is enqueued (since a sequential read may be in progress). The parameter n is preferably set comparable to or greater than the number of maximum independent activities which may be underway. Thus if any one thread is doing a sequential read the controller will perform readahead; but if all accesses are purely random, the controller will almost never do a lookahead read.

Periodic Activation of Physical Drives

If a disk drive fails in service, the user wants to know about it. However, in a system with composite disk drives, the host system may not detect the failure status unless it happens to request an access which requires access to the failed drive Even if the monitor software on the host system periodically queries the drive controller (through the normal high-level interface), such a drive failure will not necessarily be detected.

This invention provides an improved way for failure status to be propagated upward.

The DDA controller sends a recount command to the drives every n seconds (where n is a programmable parameter). Thus, physical failure of a drive will be reliably detected by the controller within a certain maximum time period. Thus, if the monitor utility periodically polls the controller every m seconds, failures will almost always be detected within m+n seconds.

Majority Voting to Select Among Valid Records

A very common problem is that a disk may fail, and then come up active at the next power-up. Thus, the controller may see inconsistent data on several drives, each of which claims to be valid.

The system of the presently preferred embodiment adds a time-dependent "whim" number to the configuration data (validation timestamp) in each drive. The drives which have the same generation of data should all have the same whim number. A zombie drive (one which has failed and revived) may report itself good, but it will have a whim number inconsistent with the other drives. Majority voting is used to select among the whim numbers. Thus, even if a zombie drive reports a later timestamp, it will be outvoted by the consensus of the other drives, and can then be excluded (and rebuilt).

Physical Mode

The system of the presently preferred embodiment (the Dell Drive Array, or "DDA") presents to a host operating system disk drive abstractions called Composite Disk Drives. Composite Disk Drives support logical requests that range from "read" and "write" to "scatter read with holes". A Composite Disk Drive (CDD) is physically implemented using (besides the hardware and firmware in the controller) one or more physical disk drives. As an abstraction, the CDD hides many aspects of the physical drives from the host operating system. In particular, the abstraction hides, and hence prevents access to, the actual physical disk drive hardware interface.

For setup, maintenance, and diagnostic purposes, there is a need to get closer to the physical disk drive interface than is allowed through the CDD abstraction. For example, when a physical disk drive (PDD) is "new" and not yet part of a CDD, a means is needed to test the PDD and to write configuration information onto the PDD. Even when a PDD is a part of a CDD, there is a need to test the PDD and perhaps write new configuration information onto the PDD. In addition to these straightforward needs, it turns out that there is a need to access the PDD interface in order to perform drive vendor specific functions. Since these functions are vendor specific and since vendors and these functions change over time, there is strong motivation to accommodate access to these functions without changing DDA firmware.

To meet these needs, DDA has a Physical Mode Programming Interface. This interface is not normally disclosed to DDA owners or users but is used by Dell's DDADIAG program and by the DDA specific portion of the EISA Configuration program.

Synchronization of Phvsical Mode (PM) Commands with Logical (CDD) Commands

Physical Mode commands may be issued by the host at any time, including periods where the host is also issuing logical CDD commands. PM commands must be able to run without disturbing (other than the obvious requirement to be running only one command per drive at a time) the operation of the CDD. When a PM command is received, PM checks to see if the physical drive specified is part of a CDD. If it is not, the command is run without regard to CDD interference. If the physical disk drive specified is part of a CDD, PM synchronizes the command with the CDD driver by submitting a PHYSICAL logical request to the CDD driver. When the PHYSICAL request reaches the head of the CDD request queue, the CDD driver "executes" it. Execution of the PHYSICAL command consists primarily and simply of calling the request's Return Function, which in this case, happens to be the core PM request driver. In other words, PM gets the CDD driver to run the PM command. Synchronization is obviously ensured.

In addition to the simple single command PM/CDD synchronization above, there is a multi-command synchronization mechanism that is part of and used with the primitive PM command set. When the host wants to run only PM commands on a disk drive for a period of greater than one command or wants to use the other primitive commands, the host will issue the BE_GIN_PHYS_MODE_ONLY command. When the host is ready to allow CDD commands to resume, it issues the END_PHYS_MODE_ONLY command.

When PM receives the BEGIN_PHYS_MODE_ONLY command, as with other commands, it checks to see if the physical drive specified is part of a CDD. As with other commands, if the drive specified is not part of a CDD, the command is run directly. It the drive specified is part of a CDD, PM gets the CDD driver to run the command as a PHYSICAL command return function. BPMO increments a phys_mode_only counter associated with the physical drive. It also increments a phys_mode_only counter associated with the CDD. Synchronization is attained by having the CDD driver refuse to run any commands when its phys_mode_only counter is non-zero.

Since the CDD driver will refuse to run any commands when in phys_mode_only mode, a refinement needs to be mentioned here. When PM runs a command, it does so by handing it to the CDD driver only if 1) the specified drive is part of a CDD and 2) the associated CDD is not in phys_mode_only mode.

When PM receives the END_PHYS_MODE_ONLY command, it performs the same CDD checks as with other commands. In normal operation, the associated CDD, if any, will be in phys mode only mode and so the EPMO command will be run directly by PM. The EPMO command decrements the phys_mode_only counter associated with the specified physical drive and decrements the phys mode-only counter associated with the associated CDD, if any. If the EPMO command causes a CDD's phys_mode_only counter to go to zero, the CDD is obviously no longer in phys_mode_only mode. At this point, the CDD driver may have logical CDD commands queued that it has refused to run. To ensure a timely restart of the CDD, PM must issue an innocuous command to the CDD driver but only after it is no longer in phys_mode_only mode. PM does this by issuing a PHYSICAL command to the CDD driver with an innocuous return function. The technical term for this action is "Thump", as in PM "thumps" the CDD driver.

Rudimentary PM Command Set

PM's rudimentary command set consists of a number of AT task file "like" commands that are still abstract like logical CDD commands in that they do not provide for direct access to the physical disk drive interface but are closer to the physical disk drive interface nevertheless. These commands are: READ, READ with no retries, READLONG, READLONG with no retries, IDENTIFY, READBUF, WRITE, WRITE with no retries, WRITELONG, WRITELONG with no retries, FORMAT, WRITEBUF, SETBUF, SEEK, RECAL, VERIFY, VERIFY with no retries, INIT, DIAG, READMULT, WRITEMULT, SETMULT, and RESET.

Primitive PM Command Set

PM's six primitive commands are provided through a rudimentary "EXTENDED" command, although that is an arbitrary implementation detail. The BPMO and EPMO primitive commands have already been discussed. The remaining four primitive commands provide the host almost direct contact with the physical disk drive hardware interface. The ISSUE_CMD command writes host specified values to the physical disk drive's task file registers. The RETURN_STATUS command reads the physical disk drive's task file registers and returns the values to the host. The READ_DATA command reads the indicated amount of data from the disk drive's data register and returns the data to the host. The WRITE_DATA command writes the host provided data to the disk drive's data register.

With these four primitive commands, the host can perform almost all standard task file commands, and can perform any of the vendor unique commands that we are currently aware of. Standard commands that cannot be performed include READLONG and WRITELONG. In addition, access is not provided to the alternate status register, the drive address register or to the device control register and drive interrupts are not reflected to the host. These limitations could be overcome by adding primitive commands and should not be thought of as limiting the scope of this disclosure.

The PM commands are currently used to 1) enable spindle sync on the CONNER 3204F 200 Meg drives and 2) to download firmware to the Maxtor LXT series drives.

Interface Co-Residence

When the emulating (AHA) interface is active, the native interface is also active (see FIG. 7). This capability is used to permit the monitor utility (running on the host computer) to use calls to the native mode of the array controller, while routine access requests use the emulated interface.

It should be noted that there are two types of host software used: the device driver does the minimal basic interface; the MONITOR utility is the facility which allows user to see smart info inside array. For example, the disclosed drive array controller may be used with monitor utilities which run on Novell or UNIX (or other) operating systems.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be described with reference to the accompanying drawings, wherein:

FIG. 1 is a functional/structural block circuit diagram of the first preferred embodiment disk drive array;

FIG. 2 illustrates disk drive platter synchronization;

FIGS. 3–4 are layout and schematic diagrams of a second preferred embodiment;

FIG. **5** is a flowchart for a small size write operation;

FIG. **6** is a flowchart for rebuilding lost data;

FIGS. **7–9** illustrate the layering and sequencing of commands;

FIGS. **10–11** are flowcharts showing indeterminate number fork/join process threads by the first preferred embodiment;

FIGS. **12–13** illustrate scatter/scatter; and

FIGS. **14–15** illustrate the use of logical and physical defect lists.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Generic Embodiment

FIG. **1** is a functional block circuit diagram of the first preferred embodiment disk drive array controller, generally denoted by reference numeral **100**, which includes microcontroller CPU **102** with embedded ROM **104** and RAM **106**, bus interface **108**, and five connected disk drives **110–114**. ROM **104** and RAM **106** may be separate from the CPU, and ROM **104** contains the firmware for controller **100**. System bus **109** provides a communication link between controller **100** and a host computer which uses the array of disk drives **110–114** as secondary memory. Controller **100** plus disk drives **110–114** may be physically located within the chassis of the host computer, or the controller and disk drives may have a separately cabinet connected to the system bus. Each of disk drives **110–114** is somewhat intelligent, such as having an IDE (ATA) interface plus an on-board data separator, buffer, and disk controller; and each disk drive may hold 200 MB of formatted data so that the array appears as a single 1,000 MB memory to the host. Communication from controller **100** to each of disk drives **110–114** is a simple 40 line cable with signal assignments mostly analogous to those of an AT I/O channel. With redundant storage across the disk drives, controller **100** plus disk drives **110–114** form a RAID, and the capacity apparent to the host decreases due to the redundancy.

FIG. **3** shows the layout of second preferred embodiment controller **200** for a disk drive array, and FIG. **4** illustrates controller **200** plus an array of disk drives (collectively referred to as DDA) within a computer system. Further details of controller **200** and DDA appear in the following.

Overview of Controller Operation

Controller **100** operates as follows. First, the disk drives **110–114** have their spinning platters synchronized by communication on cable **116**; FIG. **2** schematically illustrates synchronization in that sectors 120–124 on the platters of disk drives **110–114**, respectively, will pass their corresponding read/write heads **130–134** simultaneously. Platters typically spin at 3,600 rpm, so one revolution takes about 16.7 milliseconds. Each sector contains 512 bytes (½ KB) of information, and the number of physical sectors forming a single circular track around the platter may be about 40 for each of about 1,000 concentric tracks on the platter. This translates to a read/write rate of about 1 million bytes per second. (Typical disk drives include multiple platters and multiple heads, for increased capacity.)

Disk drives **110–114** each appears as a logical disk drive to controller **100** that simply has a number of cylinders (corresponding track(s) on platter(s) of a single disk drive) with a number of sectors per track, and the disk drive's controller translates the actual physical implementation to such logical appearance. Indeed, if defective sectors were discovered during manufacturer's testing, then these are remapped by the disk drive's on-board controller and not observed by communication through the IDE interface. Contrarily, the host sees the array as a single sequence of sectors numbered serially beginning at sector 0 and continuing through approximately sector 2,000,000 in the 1,000 MB nonredundancy version.

When the host sends a read or write request to the array on bus **109**, interface **108** translates the request to controller **100** which generates access requests for each of the individual disk drives **110–114**. Controller **100** sends logical commands to a disk drive, and the individual disk drive's controller does the actual generation of pulses to move the read/write head. Data read from or to be written to a platter is held in the disk drive's buffer RAM. The data read or to be written transfers between RAM **106** and buffer RAM in each disk drive; thus the disk drive delay in waiting for the read/write head to seek the correct track and the latency of the appropriate sector to spin around to the read/write head does not affect the data transfer. In fact, the delay averages 20 milliseconds, and multiple further read/write requests could be received by controller **100** during this time interval. Controller **100** maintains a queue of pending requests and may modify the commands to increase efficiency in the disk drive operation as explained in detail below; however, each disk drive **110–114** handles requests serially and completes a request prior to starting the next.

Data Striping

Controller **100** writes data to the disk drives **110–114** with data striping and parity: each 2 KB of data is split into four ½ KB portions which are written to four corresponding sectors on disk drives **110–114** (e.g., four of the five sectors 120–124) and the fifth corresponding sector is written with the exclusive OR (bitwise parity) of the data on the four sectors. Thus if any one of the disk drives **110–114** fails or develops a defect in its one of the sectors 120–124, the data can be reconstructed from the data in the other four sectors by CPU **102**; this constitutes the RAID aspect of controller **100** plus disk drives **110–114**. The data reconstruction permits continued operation even with a failed disk drive, and once a failed disk drive has been replaced, reconstruction of the data for writing onto the replacement disk drive could be performed in background without any down time. Of course, each sector will have its own CRC (cyclic redundancy code) or ECC (error correction code) bytes to check for errors in the 512 bytes of data in the sector; this error checking is in addition to the redundancy of the parity sector for the stripe of four sectors. Further, contiguous data could be written into stripes using every other sector on a track (2-to-1 interleaving); the time during which an intervening sector passes under the head can be used for loadinglunloading the track buffer and for error checking of the data just read from a sector and before the data from the sector following the intervening sector is written or read.

Small Writes

Controller **100** writes data with size smaller than a complete stripe (2 KB plus redundancy) in accordance with the flowchart in FIG. **5**. In particular, a write of data less than 2 KB will not fill up all of the five sectors 120–124, so the parity sector information must be computed from both the new data being written plus the old data in the sectors not being overwritten. For example, if sector 120 is the parity sector and if new data is being written to sectors 121–123,

then the old data in sector 124 will be needed in addition to the new data in order to compute the new parity information. Thus before the new parity information can be written to sector 120, controller **100** must perform a read of the data in sector 124 and then compute the exclusive OR of the new data plus the sector 124 data and write the result to sector 120 as the new parity information. See the lefthand portion of FIG. **5**. In short, the computation is:

$$\text{New 120}=\text{New 121} \oplus \text{New 122} \oplus \text{New 123} \oplus \text{Old 124} \quad (1)$$

Contrarily, if the new data being written will only occupy sector 121, then controller **100** reads the old data in sector 121 prior to the write of the new data and also reads the old parity in sector 120, and then controller **100** computes the new parity by exclusive OR of the old parity, the old data in sector 121, and the new data to go into sector 121. See the righthand portion of FIG. **5**; the computation is:

$$\text{New 120}=\text{Old 120} \oplus \text{Old 121} \oplus \text{New 121} \quad (2)$$

This computation gives the same result as if the analog of equation (1) were used where all of the old data to be preserved were read and used in the exclusive OR; that is, if the computation were:

$$\text{New 120}=\text{New 121} \oplus \text{Old 122} \oplus \text{Old 123} \oplus \text{Old 124} \quad (1')$$

The sameness of the results in equations (2) and (1') follows from the definition of the parity. But the use of equation (2) for the computation of the new parity implies that only two reads of old data are used, rather than the three reads of equation (1'). Also, the reads in equation (2) are on disk drives **110** and **111** which are also the disk drives that will be written; thus disk drives **112–114** are free to be separately read or written (in other tracks) simultaneously. In contrast, equation (1') will have reads of disk drives **112–114** and writes of disk drives **110–111**; that is, all five disk drives are active. Equation (2) also has an analog which could be used in place of equation (1) for the case of writing new data to sectors 121–123, namely:

$$\text{New 120}=\text{Old 120} \oplus \text{Old 121} \oplus \text{New 121} \oplus \ldots \oplus \text{Old 123} \oplus \text{New 123} \quad (2')$$

But the computation of equation (2') requires four reads and involves four of the five disk drives for both a read and a subsequent write; whereas, the computation of equation (1) only has one read but involves all five disk drives. Hence, controller **100** minimizes the number of reads for such small size writes by effectively picking the smaller number of disk drives to read or write.

In general, if an array includes N+1 disk drives with each stripe including 1 block (with a block being an integral number of sectors) of redundancy data on 1 disk drive and N blocks of data of 1 block on each of the other N disk drives, and if a write changes K of these N blocks, then for K greater than (N-1)/2 the write is not small and the N-K nonwritten blocks are read as in (1) but for K less than or equal to (N-1)/2 the number of blocks to be written is small enough to take advantage of using the old parity information and reading the old data as in (2) instead of all of the reads needed as in (1').

Data Guarding

When running as a drive array controller, controller **200** and DDA use sector interleaving or striping to increase the data bandwidth. Sector interleaving is basic to the redundancy algorithm. As each block of data is being written to the data drives in the array, microcode generates the byte by byte XOR across all bytes of the interleaved data sectors. The parity bytes are all combined into one parity sector and written onto the parity drive. The parity byte is generated by:

$$\text{parity}_{fourdrives}=\text{data3} \oplus \text{data2} \oplus \text{data1} \oplus \text{data0}$$
$$\text{parity}_{threedrives}=\text{data2} \oplus \text{data1} \oplus \text{data0}$$
$$\text{parity}_{twodrives}=\text{data1} \oplus \text{data0}$$

The parity sector is written onto the parity drive as soon as possible after the writes to the data drives. The performance overhead is simply the time required to generate the XOR of the data bytes, and the time to send the write data to the parity drive. In the case of a write of a number of sectors which does not divide evenly into the number of data drives, there is an additional overhead to read the other sectors in the same stripe, to perform a read modify write cycle to get the parity byte correct. This is not a major concern as the majority of disk operations are reads.

DDA can continue responding to system requests after a drive fails (providing the failure is not catastrophic enough to cause the entire system to fail, such as shorting power and ground). After DDA has determined that a drive has failed, it combines bytes from the remaining three data drives with the corresponding byte from the parity drive to regenerate the failed drive's data. These are:

$$\text{data0}_{fourdrives}=\text{data3} \oplus \text{data2} \oplus \text{data1} \oplus \text{parity}$$
$$\text{data0}_{threedrives}=\text{data2} \oplus \text{data1} \oplus \text{parity}$$
$$\text{data0}_{twodrives}=\text{data1} \oplus \text{parity}$$
$$\text{data1}_{fourdrives}=\text{data3} \oplus \text{data2} \oplus \text{data0} \oplus \text{parity}$$
$$\text{data1}_{threedrives}=\text{data2} \oplus \text{data0} \oplus \text{parity}$$
$$\text{data1}_{twodrives}=\text{data0} \oplus \text{parity}$$
$$\text{data2}_{fourdrives}=\text{data3} \oplus \text{data1} \oplus \text{data0} \oplus \text{parity}$$
$$\text{data2}_{threedrives}=\text{data1} \oplus \text{data0} \oplus \text{parity}$$
$$\text{data3}_{fourdrives}=\text{data2} \oplus \text{data1} \oplus \text{data0} \oplus \text{parity}$$

This is done by microcode and data accesses suffer a slight performance degradation after a drive has failed.

The reason that only 8 bits of data are required for complete redundancy is that we know which 8 bits are wrong. Most of the bits in a typical ECC design are used to find the location of the erroneous data bits. In this case there is a CRC or ECC field as part of every sector on the disk drive. The drive controller chip checks the CRC or ECC, and reports whenever there is erroneous data. Since each drive controller chip handles a different byte of the data, we can determine which byte is bad.

Data guarding suffers a fairly severe write performance degradation as partial stripe write must do a read modify write cycle to correctly update the parity data, and this will usually result in a lost revolution of the disk.

Mirroring

Mirroring is a data redundancy technique in which two complete sets of the data are kept. DDA supports this in hardware, or more appropriately in firmware. When a drive fails in mirrored mode, the data is simply recovered from the second copy on the second set of drives.

Mirroring has much better performance than data guarding, but it requires a higher overhead in disk drives.

Rebuilding

When the number of defective sectors has grown so large as to exhaust the capacity of one of the disk drives in an array, then that disk drive must be replaced in order to continue complete data guarding. Similarly, if a disk drive fails for some other reason such as mechanical failures

including head crashes, then the disk drive must be replaced for continued data guarding. A replacement disk drive must be loaded with the data from the failed disk drive, and controller **100** accomplishes this by use of the redundancy: each sector of data in the failed disk drive can be reconstructed using the corresponding data and parity sectors of the still-viable disk drive. In essence, controller **100** rebuilds the data of the failed disk drive into the replacement disk drive by reconstructing the data bit by bit from the remaining data and parity. Of course, this also applies in the same manner when a parity disk drive fails.

After installation of a replacement disk drive, controller **100** uses its defect physical location table to unmap all of the remapped sectors which were originally remapped due to defective sectors on the replaced disk drive; this should release spare sectors for remapping of future developing defects, especially if the disk drive with the greatest number of defects had been replaced. Of course, replacement of a disk drive requires controller **100** to reconstruct the data lost by use of the redundancy/parity of the other disk drives, and controller **100** must also systematically rebuild the data lost on the new disk drive. Controller **100** maintains two independent representations of all drive defects. The first defect list is used to maintain the sector remapping structure when reconstructing redundancy information. This list is called the logical defect list and is stored in the remap data structure. The second list, called the physical defect list, is used to preserve known defect information on a physical disk basis. It allows defects to be maintained across logical configurations and is stored in a special reserved area which is always known even without a logical drive configuration. Controller **100** uses both the physical defect list and the logical defect (remap) list for this unmapping.

Controller **100** rebuilds the data lost on a replacement disk drive according to the method indicated by the flow chart in FIG. **6**; this allows continued access to the disk drives and rebuilding in background unless access density precludes timely rebuilding. As shown by the righthand branch in FIG. **6**, controller **100** effectively waits to begin rebuilding until no access to the disk drives has been requested for an interval of 125 milliseconds; controller **100** interprets such a no request interval as the start of a request hiatus. At the end of the 125 millisecond interval controller **100** begins rebuilding until it receives an access request. Of course, controller **100** rebuilds stripes sequentially to use minimal read/write head movement. Note that if rebuilding were to be interleaved with access request servicing, then the repetitious read/write head seeking tracks between the rebuilding regions and the accessed regions result in disk thrashing and decreased throughput; thus the 125 millisecond wait is an indication of a request hiatus. However, as shown in the lefthand branch of FIG. **6**, controller **100** guarantees the rebuilding of the data lost onto the new disk within a reasonable time by the provision that if no 125 millisecond hiatus has occurred for 750 milliseconds, then controller **100** inserts a fixed amount of rebuilding (e.g. all sectors on one track) despite pending requests for access.

The parameters of the rebuilding may be adjusted for the type of use anticipated; that is, the 125 millisecond wait may be lengthened or shortened, the 750 millisecond interval may be adjusted, and the fixed amount of inserted rebuilding may be varied.

Dual Defect Lists

Note that, as the foregoing description specifies, TWO independent representations of drive defects are preferably maintained. The first defect list is used to maintain the sector remlapping structure when reconstructing redundancy information. This list is called the logical defect list and is stored in the remap data structure. The second list, called the physical defect list, is used to preserve known defect information on a physical disk basis. It allows defects to be maintained across logical configurations and is stored in a special reserved area which is always known even without a logical drive configuration.

Periodic Activation of Drives

If a drive fails in service, the user wants to know about it. However, unless the host system happens to request an access which requires access to the failed drive, it may not detect the failure status. Even if the monitor software on the host system periodically queries the drive controller (through the normal high-level interface), such a drive failure will not necessarily be detected.

The DDA has a variable for tracking requests to the drives, and if no requests are enqueued for a period of n seconds (where n may be a programmable parameter and typically equals 1), then DDA sends a command (such as a recount command) to the drives to check their viability. Thus, physical failure of a drive will be reliably detected by the controller within a certain maximum time period. So if the monitor utility periodically polls the controller every m seconds, failures will almost always be detected within m+n seconds. In the event that the controller has partitioned the physical disks into two or more groups with each group appearing as a disk drive or a disk drive array to the host, then a tracking variable will be established for each group.

Drive Data Restoration

When DDA detects a drive failure, it first determines whether it is due to a localized media defect (a so-called grown defect), or whether the entire drive has failed. In the case of a grown defect, the bad block or blocks are remapped to another portion of the disk as discussed below.

When DDA determines that a drive has failed completely, it marks the drive as bad and never attempts to access it again. This information is stored on all the other disks, and so it is maintained across power up cycles and resets. If one of the data redundancy modes is in effect on the logical disk with the failed drive, DDA will continue to operate recovering the data as described above.

DDA has no way of informing the user that a drive has failed, and so host software must poll DDA on occasion to determine this and report it to the user in an appropriate fashion. This is covered in more detail in the following.

On power up, DDA determines which drives have failed, which drives have failed previously and now seem OK, and whether a drive has been replaced. DDA can differentiate between new drives and previously bad drives, by looking at the DDA sector (see above); new drives are assumed not to have a DDA sector. To reuse a drive which has previously failed, the diagnostics must be used to erase the drive's DDA sector.

After the failed drive has been replaced, DDA will rebuild the failed drive's data on the new drive with the algorithms outlined above. This can occur in either a foreground or background mode depending on whether the user needs to use the system in a hurry. The appendix firmware implements this rebuilding analogous to the first preferred embodiment's rebuilding as illustrated in FIG. **6**.

Bad Block Remapping

DDA always presents perfect defect free disks to the host. The drives are supposed to come from their manufacturer

defect free, but DDA allows for a slight relaxation of this and for grown defects which express themselves after manufacturing test.

Bad block mapping only works for logical drives with redundancy, because when DDA moves the bad block, it needs to be able to reconstruct the data. Bad block remapping is performed on stripe boundaries; if any block in a stripe is bad, the entire stripe is remapped.

Referring now to FIG. **14**, remapping occurs dynamically analogous to the procedure described in connection with the first preferred embodiment. When DDA determines that a block is bad, it performs the following procedure:

1. Allocate space for the stripe in the alternate block reserved section of the disk (see above).
2. Recover the data in the stripe containing the bad block.
3. Write the recovered data to the newly allocated stripe.
4. Update the remap tables in memory with the remap information and an indication of which drive or drives failed.
5. Write the remap tables to each disk in the logical drive.
6. Retry the operation which originally failed and uncovered the bad block.

Referring now to FIG. **15**, in normal operation, on every disk access DDA searches the remap tables to determine if the requested access will intersect one or more of the bad blocks on the disk. If so, the access is split into multiple accesses in which the accesses to remapped blocks are remapped to their alternate sectors. This will incur a performance degradation on remapped blocks because of the extra seeks required. The remap table search will not cause any noticeable performance degradation.

The number of bad blocks which can be handled is configurable through the EISA configurator (see above).

Non-redundant Remapping Options

Alternatively, DDA could remap blocks on non-redundant modes when the drive reports a correctable error. That is, blocks which contain defects but with the CRC or ECC providing sufficient information to overcome the defects (i.e., soft errors) may be remapped. Soft errors are transparent to controller **100**, but their locations can be made available.

Second Embodiment

FIG. **3** shows a layout on a card for second preferred embodiment controller **200** for a redundant array of disk drives (DDA), and FIG. **4** schematically shows controller **200** incorporated in a system with an EISA bus and ten IDE (ATA) disk drives. Controller **200** may be termed a DDA adapter and consists of five major functional blocks: (1) Local Processor, (2) EISA Bus Interface including BMIC, (3) AHA register emulation, (4) IDE Drive Interface, and (5) an optional Cache.

Local Processor

The local processor is denoted by reference numeral **208** and is an Intel 80960KA (a 32-bit RISC microcontroller) running at 16 MHz and has local 512K of ROM **209** and 256K SRAM **215**. ROM **209** holds the microcode for the base functionality, and a monitor allows microcode to be loaded into local RAM **215** and then executed. Appendix A contains a listing of the firmware. Static RAM **215** is implemented in eight 32Kx8 SRAM modules. These modules are connected as two interleaved banks of 32Kx32. The interleave makes it possible for the 80960 to access the SRAM in 3-1-1-1 clocks, which results in one wait state for the first access of a burst and zero wait states for subsequent accesses. The 80960 directly supports four interrupts. These will be used for:

BMIC interrupt.

AHA-1540 register interrupt.

A logical OR of the ten drive interrupts.

A timer tick interrupt.

EISA Interface (BMIC Chip)

The EISA bus interface is handled mostly by an Intel 82355 BMIC bus master interface chip denoted by reference numeral **211**. BMIC handles local processor **208** accesses to system memory through peek and poke registers. It also has a two channel DMA controller which performs EISA bursts at full bus bandwidth.

The DMA controller in BMIC does not provide addressing for the transfer. BMIC in controller **200** talks to a DMA controller which increments the addresses for the static RAM accesses. During BMIC transfers the DMA controller takes complete control of the system bus of 80960, and 80960 is put into hold for the duration of the transfer.

BMIC only has 128K addressability, so there is a DMA page register bit to allow transfers from all of the static RAM. The DMA controller has an approximate maximum bandwidth of 28 Megabytes/sec which is below the EISA maximum but sufficient for controller **200**.

AHA Emulation

AHA-1540 emulation is provided by dual porting the static RAM. All the AHA registers except the data register are mapped to specific locations in the static RAM. The majority of the AHA register semantics are handled in microcode. There are however some exceptions:

Writing to any AHA register generates an interrupt.

Reading the AHA status register generates an interrupt.

The AHA status register bits CDF and DF are implemented in hardware and forced onto the EISA data bus on reads from that register.

IDE Interface

FIG. **3** also shows IDE disk drive connectors **201–205**, bus connector **210** for a 32-bit EISA bus (157 lines), daughter card connector **212** which would connect to an optional cache daughter card, and IDE drive interface chips labelled "IDE 1" through "IDE 5" and located adjacent the drive connectors **201–205**. The IDE disk drives may be models such as the Conner CP3204.

Bibliography

The following references contain further information about elements of controller **200** and are hereby incorporated by reference:

Intel 80960KB Programmers Reference Manual

Intel 80960KB Hardware Reference Manual

Intel 80960KA Data Sheet

Intel 82355 Bus Master Interface Chip Data Sheet

Committee Draft ATA drive interface proposal

EISA Specification Version 3.10

Conner CP3204 Product Manual

Adaptec AHA-1540B/1542B User's Manual.

Features

Controller **200**/DDA adapter has the following features:

Support for DOS, OS/2, Lan Manager, Novell, and UNIX.

Modular design consists of a base drive controller card plus an optional daughter card which provides for a disk cache.

Controls multiple parallel arrays of 2,3,4, or 5 IDE drives each giving a corresponding increase in disk transfer rate.

In a parallel array one of the drives can be used for parity data giving complete redundancy given a single drive failure.

After a failed drive in a redundant array has been replaced, its data can be rebuilt.

Array supports sector interleave.

After a drive has failed the array will collect data on the fly, and continue operation with a slight performance degradation.

Supports synced and non-synced spindle standard off the shelf IDE drives.

Optionally controls the same number of independent drives with parallel, overlapped seeks and data transfers on up to ten drives at one time.

Supports firmware mirroring of drives.

Supports EISA bus master protocols including burst transfers.

Optional one to thirty-two megabyte disk cache on daughterboard.

Patchable microcode

Compatible with the Adaptec SCSI bus master register sets so that the Adaptec drivers will operate the card.

Support for multiple controllers.

Support for coresidence with IDE, ESDI, and SCSI host adapters.

Cables

Included with the DDA adapter itself are two types of cables, the "super cable" assembly and the "sync cable". The super cable is the set of five data cables which connect the disk drives to the adapter board. The sync cable is used to communicate between the drives themselves, so that they may synchronize their rotational speed and position.

Modes of Operation

The DDA adapter operates in one of two modes. There is a native interface which is always active and is used by the utility software for testing and configuration purposes. This interface is also used to support the BIOS interface and native device drivers. In this mode, the option ROM BIOS on the adapter board provides BIOS interrupt 13h compatibility for traditional MS-DOS applications. The other operating mode, chosen to provide broad compatibility with industry-standard operating system software, is the Adaptec AHA emulation mode. In this mode, the DDA adapter emulates the Adaptec AHA-154x series of SCSI disk adapter boards. Both operating modes are implemented in the firmware supplied with the DDA adapter.

EISA Configuration Utility

An EISA utility allows the user to fully configure the DDA adapter. There are no physical jumpers on the DDA adapter board itself.

Firmware

There are two distinct areas of firmware on the DDA adapter. First, there is the Option ROM BIOS, which is contained in the DIP EPROM near the bus connector. This contains code which is executed by the host CPU on power-up. The second area is contained in the four burst mode EPROMs on the other end of the DDA adapter board. These contain the working code for the 80960KA CPU. This code provides the operating modes, reads and writes the disk drives, manages the RAM cache and sets up the bus master mode DMA over the EISA bus. All of the EPROMs are socketed and may be removed and replaced.

Firmware upgrades may thus take three forms. First, the Option ROM BIOS EPROM may be exchanged with one containing updated object code. Second, the four burst mode EPROMs may be exchanged. Third, there is a hot patch mechanism which allows minor code changes to be applied via a self-booting diskette. This patch technique stores new 80960 object code on the disk drives themselves, and the existing firmware reads this into RAM at power-up time and links it with itself. There is a limited amount of RAM on the adapter board, so that this technique is limited to minor upgrades.

The DDA firmware is divided into functional groups and each group further divided into successively smaller subgroups until each firmware function is sufficiently small in both scope and code size. References to these firmware functions are made exclusively through function pointers rather than direct calls. Using this technique, and firmware function may be entirely replaced without disrupting the main body of the firmware. Where a typical firmware initialization would consist of a single general initialization, DDA's firmware initializes in two distinct phases; first the public function pointers, known as patch pointers, are established, and then general initialization itself is fully patchable. (Each functional group is responsible for implementing its own patch and initialization phases.) A typical patch phase would look similar to the following.

```
emulpatch()
{
    emulISR = patch(DefISR);
    emulDecode = patch(DefDecode);
    emulReadAhead = patch(DefReadAhead);
    emulInit = patch(DefInit);
}
```

In this example, the Def* symbols represent function entry points found in the DDA ROM and the emul* symbols represent patch pointers. The patch( ) function returns a pointer to a function, either the pointer passed to it or a RAM-based replacement function. Code which desires to use any of the emul* functions in this example would call the function(s) through the patch pointers and would not know whether the actual service routine is running from ROM or from RAM.

Downloading firmware may contain the following items: code (.text objects), constant data (data in the text segment), initialized data (.data objects), and uninitialized data (.bss objects). Objects in the .bss section will be initially zero-valued. Patch routines (downloaded firmware) may not call ROM routines directly, but must do so through patch pointers. If such a pointer does not exist, one can be created and placed in the .text or .data segments. Patch routines cannot know the addresses of objects in other patch routines. For this reason, a generic array of integers is reserved in the ROM data segment. This area can be used to allow several patch routines to share data.

At the beginning of the patch initialization phase, a firmware data structure is loaded from disk. This data structure is known as the patch record and is described below:

Field	Size
DDA ROM target version	4 bytes
patch version number	4 bytes
patch CRC	4 bytes
total length	4 bytes
patch block 1	variable
patch block 2	variable
patch block 3	variable
...	variable

The DDA ROM target version must match the actual ROM version. This field exists so that DDA can automatically invalidate the patch record when its firmware ROMs are upgraded.

The patch version number is a 32-bit integer providing a unique identifier for the patch record. This field is made available to host software after DDA initialization.

The patch CRC is used by the firmware to verify the correctness of the patch record.

The patch length is used to calculate the actual patch CRC.

Firmware Overview

The DDA firmware consists of approximately 40,000 lines of C and 80960KA assembly and uses an interrupt-driven, non-tasking model. The firmware consists of several functional types: interrupt service routines (ISRs), request decoders, global resource managers, resume functions, and device drivers. Interrupt service routines (ISRs) are functions that are responsible for handling the occurrence of a hardware event (interrupt). Request decoders are functions that convert generic I/O requests into device driver requests and allocate global resources for those device requests. Global resource managers are modules that manage the allocation and deallocation of global resources. Resume functions are functions that are called upon completion of a device request by a device driver. These functions are responsible for the control flow and error reporting. Device drivers are modules that manage hardware devices as global resources. Upon receipt of an I/O request from the host, the firmware initially processes the request at user (non-interrupt) privilege level. once the request is decomposed into manageable pieces and enqueued on either the disk driver's queue (reads) or the transfer driver's queue (writes and cache hit reads), the first hardware command is launched. The hardware operation ultimately generates an interrupt to the processor. All other request processing generally occurs in interrupt service routines.

The DDA firmware is composed of three conceptual layers: the host interface layer, the request processor layer, and the device driver layer (see FIG. 7). The first conceptual layer, the host interface, is made up of the native interface, the AHA interface, and a set of internal initiators. The host interface layer is responsible for initiating all I/O activity in the controller and its three components can be active simultaneously. This coactivity allows the DDA firmware to perform such things as background event logging and rebuilding, but more interestingly, it allows DDA to emulate non-intelligent controllers while simultaneously supporting array monitoring facilities through its native interface. This design does introduce the complication that host software may believe there are two controllers in the system and try to install two device drivers, so DDA's emulation mode is defeatable and a command is provided in the native interface to allow native device drivers to check the emulation mode.

The second layer, the request processor, consists of a set of request decoders, resume functions, and global resource management functions. The request decoders are responsible for decomposing I/O requests into manageable pieces, allocating global resources such as disk buffers and request structures, and submitting requests to the disk and transfer drivers. All request decomposition occurs at user level using a single thread of execution. Because of this, all functions responsible for request decomposition are allowed to be non-reentrant. This greatly simplifies the firmware that manages allocation of global resources. In contrast, the resume functions all run at interrupt privilege. Because of this, resume functions cannot be interrupted by hardware events of the same type associated with the resume function. All global resource management functions, both allocation and deallocation, are non-reentrant. The DDA firmware uses a policy where all global resources are allocated at user privilege by single threaded code and all global resources are deallocated at maximum privilege level. The purpose of making all global resource management functions non-reentrant is to minimize the number of critical regions in the firmware. Critical region implementation is very slow on the DDA hardware platform. Since deallocation functions are non-reentrant, all code calling these functions must insure that they are at maximum privilege level. This is again done for performance reasons. Under normal flow of control, callers of deallocation functions are naturally at the correct privilege level.

The third layer, the device layer, consists of three device drivers: the transfer driver, the disk driver, and the timer. Device drivers typically consist of an enqueue routine, a command processor, and an ISR. These drivers make up the heart of the DDA firmware, as they perform all of the actual data transfer and manipulation that the disk subsystem is tasked with performing. The device driver enqueue routines are fully reentrant, though the command processors and ISRs are not.

FIGS. **8** and **9** illustrate the sequence of operations in a read and write, respectively, in terms of operation privilege level. In particular, the lefthand edge of FIG. **8** indicates either a Native or an AHA request from the host as an interrupt service routine (inputISR) generated in the Native Interface or AHA Interface layer and with Xfer or Input privilege level. Execution in the Request Processor/Resource Management layer (with Usr privilege level) of the decode and allocation functions (decode and alloc) allocates buffer memory plus generates disk requests and transfer requests. The disk requests are enqueued (dskEnq) and when an enqueued disk request reaches the head of the queue, the Disk Drive layer issues a disk read command (dskCmd) with privilege level Dsk to the disk drives. When the disk drives complete the read by putting the read data in the allocated buffer memory, it issues an interrupt which is serviced (dskISR) with Dsk privilege level. This runs the disk programmed I/O function (dskPIO) still with Dsk privilege level. Then the transfer requests for buffer to host transfer are enqueued (xfer Enq). When an enqueued transfer request reaches the head of the queue, the transfer command (xferCmd) issues in the Host Transfer Driver layer and the data transfers from buffer memory to the host. Upon completion of the data transfer, the event completion interrupt is serviced (xferISR) in the to run the completion and free the allocated resources functions (complete and free). Note that

the transfer to the host has the highest privilege level, whereas the decoding and resource allocation has the lowest privilege level. This privilege hierarchy helps avoid deadlocks. The write sequence in FIG. **9** is analogous.

Specific Data Structures

Processor **208** (the 80960) creates a structure in response to the request from the host which uses the six entries in the mailbox for the first six fields:

```
typedef struct _ntvRequest
{
    uchar command; /* command byte */
    uchar drive; /* drive number */
    uchar sectorCount; /* transfer count */
    uchar handle; /* host request ID */
    union secnum sn; /* starting sector number */
    union hostaddr ha; /* host address */
    struct _ntvRequest * volatile next;
    ulong finalSector;
    volatile ushort forkCount;
    volatile ushort joinCount;
    uchar status;
    uchar remaining;
    uchar index;
    uchar tCount;
    ulong tSector;
} ntvRequest;
```

The other fields of the structure include a pointer to the next host request structure, counters for request multithreading analogous to that described in connection with FIGS. **10** and **11**, and error tracing and reporting aids. Further, processor **208** will generate corresponding structures for submission of requests to the IDE disk drives as follows:

```
typedef struct _ideRequest
{
 uchar command; /* ide logical commands */
 uchar drive;
 uchar sectorCount; /* max request size is 255 sectors */
 uchar status;
 ulong sectorNumber; /* logical sector number */
 ushort *buffer;
 void *requestPtr; /* pointer to originating request struct */
 struct _bmicRequest *bmicReq; /*pointer for completion routines */
 void (*returnFun)(); /* function to call when the operation */
    /* is complete. This function gets called */
    /* with one argument which is the pointer */
    /* to this structure. */
 struct _ideRequest *next; /* linked list pointer for queues */
 struct _ideRequest *previous;/* linked list pointer for queues */
 uint decodePhase; /* used by decoder to handle resets */
 uint retries; /* retry count for this request */
 ushort justifyBM; /* bitmap of failed drives for REMAPPING */
 uchar type; /* ORIGINAL, FRAGMENT, COMBINATION... */
 uchar combineEligible;/* TRUE/FALSE eligible for COMBINATION */
#ifdef TIMESTAMP /* Time in milliseconds when request got */
 long timeStamp; /* to this routine */
#endif
} ideRequest;
```

Also, processor **208** generates structures for transfers to and from the host's primary memory through BMIC as follows:

```
typedef struct _bmicRequest
{
 uchar command; /* use #defines above */
 char xferPriority; /* priority of transfer; values of 0 to 3 */
    /* are allowed; 0 is highest priority. */
 char disableXfer; /* disable actual transfer if asserted */
 char *radAddr; /* address of DDA side buffer */
 uint hostAddr; /* address of host side buffer */
 uint length; /* length of transfer in bytes */
 struct _bmicRequest *next; /*      */
 void *requestPtr; /* pointer to originating request struct */
 struct _ideRequest *ideReq; /* pointer used by completion routines
*/
 void (*returnFun)(); /* The function to call when the operation */
    /* is complete. This function gets called */
    /* with one argument which is a pointer */
    /* to this structure. */
 uint decodePhase; /* used by decoder to handle resets */
} bmicRequest;
```

As example, a write request from the host with controller **200** in native mode proceeds as indicated by the following simplified code which proceeds roughly as shown in FIGS. **10–11**:

```
void ntvcompleteRequest(ntvRequest *ntvReq)
{
 BMIC_INDEX(BMIC_SEMAPHORE_1)
 BMIC_SDATA(1);
 BMIC_INDEX(BMIC_LIR_AUTOINC | BMIC_MBOX_13);
 BMIC_SDATA(ntvReq->status);
 BMIC_SDATA(ntvReq->remaining);
 BMIC_SDATA(ntvReq->handle);
 BMIC_INTERRUPTHOST(NTV_LOGICAL_DOORBELL);
 ntvReq->command = NTV_NOCOMMAND;
}
void ntvIdeResumeWrite(ideRequest *ideReq)
{
 ntvRequest *ntvReq = ideReq->requestPtr;
 if (ideReq->decodePhase != ntvPhase)
 {
  disableInts();
  FreeSectorBuffer(ideReq->buffer, ideReq->sectorCount);
  FreeIdeReqBuffer(ideReq);
  enableInts();
  return;
 }
 if (ideReq->status) /* an error occurred when attempt write*/
 {
  int cnt = ntvReq->finalSector - ideReq->sectorNumber;
  if (ntvReq->remaining < cnt) ntvReq->remaining = cnt;
  ntvReq->status = ideReq->status;
 }
 disableInts(); /* start join operation */
 if ((++(ntvReq->joinCount) == ntvReq->forkCount) &&
  (ntvReq->sectorCount == 0))
 {
  ntvcompleteRequest(ntvReq);
 }
 if (ideReq->status)
  FlushSectorBuffer(ideReq->buffer, ideReq->sectorCount);
 FreeSectorbuffer(ideReq->buffer, ideReq->sectorCount);
 FreeIdeReqBuffer(ideReq);
 enableInts();
}
void ntvBmicResumeWrite(bmicRequest *bmicReq)
{
 register ideRequest *ideReq = bmicReq->ideReq;
 if (ideReq->decodePhase != ntvPhase)
 {
  FreeSectorBuffer(ideReq->buffer, ideReq->sectorCount);
  FreeIdeReqBuffer(ideReq);
  FreeBmicReqBuffer(bmicReq);
  return;
 }
 doIde(ideReq);
 FreeBmicReqBuffer(bmicReq);
}
void ntvDecode(void)
{
```

-continued

```
ntvRequest *ntvReq;
bmicRequest *bmicReq;
ideRequest *ideReq;
unsigned int thisPhase;
unsigned int count;
unsigned short *buffer;
if (ntvState & 0x00000080) /* any to decode? */
{
 thisPhase = ntvPhase;
 ntvReq = ntvRequests + (ntvState & 0x0000007F);
 switch (ntvReq->command)
 {
  case NTV_WRITE :
   count = GetWriteBuffer((int)ntvReq->drive,
    (int)ntvReq->sn.sectorNumber,
    (int)ntvReq->sectorCount,
    &buffer);
   if (count == 0) break;
   bmicReq = (bmicRequest *)GetBmicReqBuffer();
   if (bmicReq == NULL)
   {
    disableInts();
    FlushSectorBuffer(buffer, count);
    FreeSectorBuffer(buffer, count);
    enableInts();
    break;
   }
   ideReq = GetIdeReqBuffer();
   if (ideReq == NULL)
   {
    disableInts();
    FlushSectorBuffer(buffer, count);
    FreeSectorBuffer(buffer, count);
    FreeBmicReqBuffer(bmicReq);
    enableInts();
    break;
   }
   bmicReq->command = BMIC_READFROMHOST;
   bmicReq->radAddr = (char *)buffer;
   bmicReq->hostAddr = ntvReq->ha.hostAddress;
   bmicReq->length = count * BUFFERSIZEBYTES;
   bmicReq->requestPtr = ntvReq;
   bmicReq->ideReq = ideReq;
   bmicReq->decodePhase = thisPhase;
   bmicReq->next = NULL;
   ideReq->drive = ntvReq->drive;
   ideReq->sectorCount = count;
   ideReq->sectorNumber = ntvReq->sn.sectorNumber;
   ideReq->buffer = buffer;
   ideReq->requestPtr = ntvReq;
   ideReq->decodePhase = thisPhase;
   bmicReq->returnFun = ntvBmicResumeWrite;
   ideReq->returnFun = ntvIdeResumeWrite;
   ideReq->command = WRITE;
   ntvReq->forkCount += 1;
   ntvReq->sectorCount -= count;
   ntvReq->sn.sectorNumber += count;
   ntvReq->ha.hostAddress += count * BUFFERSIZEBYTES;
   if (netReq->sectorCount == 0) ntvAdvanceState(ntvReq);
   bmicBurst(bmicReq);
   break;
 }
 }
}
void main(void)
{
 /* power up initialization */
 for (;;)
 {
  if (ntvState) ntvDecode();
  /* unrelated further */
 }
}
```

The main function calls the ntvDecode function if the ntvState is nonzero which means that a request from the host is pending. The ntvRequest structure initially has the values for number of sectors and addresses in the host's primary memory of the data to be written as requested by the host.

The portion of the ntvDecode function pertinent for a write assigns the number of sectors to be written in the first submission to the disk drives to the count variable by use of the GetWriteBuffer function; typically insufficient resources such as buffers will be available to cover the entire request from the host at once, so a portion of the request will be forked off and submitted and the remainder of the request will wait for further resource availability. That is, the count variable will be typically be less than the total number of sectors to be written pursuant to the request of the host. The ntvDecode function then initializes members of the bmicRequest and ideRequest structures using the count variable for how much to read from the host and the ntvRequest structure member values for addresses, and then adjusts the member values in ntvRequest to compensate for the portion to be written; that is, the ntvDecode function increments the forkCount of ntvRequest to reflect a portion request is being submitted to the disk drives, decrements the sectorCount by the value of the count variable to keep track of the number of sectors still to be submitted, and increments the sectorNumber and hostaddress by the value of count and the value of count multiplied by sector size to keep track of the next portion's starting point in the host's primary memory, respectively. The return functions ntvBmicResumeWrite and ntvIdeResumeWrite come from the bmicRequest and ideRequest structures, respectively; the doide function in the ntvBmicResumeWrite function makes the submission to the disk drives, and the ntvIdeResumeWrite function increments the joinCount member of ntvRequest upon completion of the disk drive write. The first function, ntvCompleteRequest, is called from the ntvIdeResumeWrite function if the total write has been completed (forkCount equals joinCount, so nothing pending, and sectorCount equals zero so nothing left to write) and sets BMIC interface **211** to interrupt the host to indicate completion.

Additional Data Structures

The following code shows more actual data structures used in the presently preferred embodiment. However, those skilled in the art will of course recognize that this implementation can be very widely modified and varied.

```
/*
** types.h -- DDA and NINDY data type definitions.
*/
typedef unsigned char uchar;
typedef unsigned short ushort;
typedef unsigned int uint;
typedef unsigned long ulong;
typedef void (*pfunc)(void);
typedef struct {
 short reservedCylinders; /* rsvd tracks per physical drive */
 short heads; /* number of heads in the logical drive */
 short sectors; /* sectors per track */
 short cylinders; /* */
} geometry;
/* iac structure */
typedef struct {
unsigned short field2;
unsigned char field1;
unsigned char message_type;
unsigned int field3;
unsigned int field4;
unsigned int field5;
} iac_struct;
/*
** extracted from NINDY definitions
*/
struct file { /* file header structure */
unsigned short file_type; /* file type */
```

-continued

```
unsigned short num_secs; /* number of sections */
long time_date; /* time and date stamp */
long symtbl_ptr; /* symbol table ptr */
long num_syms; /* num entries in symb tabl */
unsigned short opt_hdr; /* size of optional header */
unsigned short flags; /* flags */
};
struct sect { /* section header structure */
char sec_name[8]; /* section name */
long p_addr; /* physical address */
long v_addr; /* virtual address */
long sec_size; /* size of sections */
long data_ptr; /* pointer to data */
long reloc_ptr; /* relocation pointer */
long line_num_ptr; /* line number pointer */
unsigned short num_reloc; /* number of reloc entries */
unsigned short num_line; /* number line num entries */
long flags; /* flags */
unsigned long sec_align; /* alignment for sect bndry */
};
struct aout { /* a.out header structure */
unsigned short magic_nmbr; /* magic number */
unsigned short version; /* version */
long text_size; /* size of .text section */
long data_size; /* size of .data section */
long bss_size; /* size of .bss section */
long start_addr; /* starting address */
long text_begin; /* start of text section */
long data_begin; /* start of data section */
};
#define FHDRSIZE sizeof(struct file) /* size of file header */
#define SHDRSIZE sizeof(struct sect) /* size of section header */
#define AOUTSIZE sizeof(struct aout) /* size of optional header */
/* The size of the above structures NOT INCLUDING any padding added
* by the compiler for alignment of entries within arrays. */
#define FHDR_UNPADDED_SIZE (3*sizeof(long) + 4*sizeof(short))
#define SHDR_UNPADDED_SIZE (8*sizeof(long) + 2*sizeof(short) +
8*sizeof(char))
#define AOUT_UNPADDED_SIZE (6*sizeof(long) + 2*sizeof(short))
/* file header union */
union _filebuf {
unsigned char buff[FHDRSIZE];
struct file filehead;
};
/* section header union */
union _sectbuf {
unsigned char buff[SHDRSIZE];
struct sect secthead;
};
/* a.out header union */
union _aoutbuf {
unsigned char buff[AOUTSIZE];
struct aout aouthead;
};
struct fault_data {
unsigned reserved;
unsigned override[3];
unsigned fdata[3];
unsigned override_data;
 unsigned pc;
unsigned ac;
unsigned int fsubtype:8,
freserved:8,
ftype: 8,
fflags:8;
unsigned int *faddress;
};
typedef struct
{
 long sector;
 ushort offset;
 ushort justifyBM;
} remapEntry;
/*
** The following is the cache block data structure.
** index is a signed short. We can have no more than 32768 buffers.
** This is not a problem since >32768 buffers would consume a massive
** amount of memory. We will have to define a buffer to be a cache
** block to support >7 megs of cache memory.
*/
```

-continued

```
typedef struct _sectorbuf
{
 struct _sectorbuf * volatile nextfree; /* next in free list */
 struct _sectorbuf * volatile nexthash; /* next in hash table*/
 int sectoraddress; /* sector address *** */
 short index; /* logical number of sectorbuf */
 char flush; /* is this buffer free? */
 char free; /* need to flush buffer? */
} sectorbuf;
typedef struct _patchHdr
{
 void *romAddr; /* ROM routine this is designed to replace */
 char *loadAddr; /* location in DDA memory where loaded */
 struct _patchHdr *next; /* pointer to next patch function */
```

Resource Limitations

In addition to fragmenting requests for remapping and combining sequential requests, controller **100** may also decompose a request when insufficient resources, such as data buffers, are currently available to satisfy the request; this permits portions of the request to be satisfied as the resources become available as illustrated in FIG. **10**. In particular, if a read request from the host requires reading 160 consecutive sectors at a time that data buffers sufficient for only 20 sectors are available, then controller **100**, following the procedure of FIG. **10**, initializes a fork counter to 0, a join counter to 0, and a remain counter to 160. Next, controller **100** allocates request structures according to the available data buffers for 20 sectors and the remaining unavailable 140 sectors. Then controller **100** allocates data buffers for the 20 sectors and forks off the request for the 20 sector read and submits this to the disk drives leaving the 140 sector read pending until further data buffers become available and another request for a portion can be forked off. Note that for efficiency considerations, CPU may avoid allocations based on small amounts of available data buffer and wait until larger amounts of data buffer become available. The net result of the procedure of FIG. **10** is a decomposition of the original host request by forking a series of smaller requests to the disk drives when data buffers become available; this forking does not have a predetermined number of smaller requests and does not require any software critical regions. The fork counter keeps track of the number of smaller requests submitted, and the remain counter keeps track of the remaining number of sectors not yet covered by a submitted request.

The individual disk drives report the completions of the smaller requests and controller **100** invokes the procedure shown in FIG. **11**. Indeed, each request completion increments the join counter, and if the join counter equals the fork counter and if the remain counter is zero, then the entire host request has been completed and the host is notified. Otherwise, each request completion just releases its data buffers and request structure. The description of the second preferred embodiment includes a more explicit description, including code, for an analog of the procedure of FIG. **11**.

Performance-Enhancement Optimizations

As shown in FIGS. **3–4**, there are connectors for up to ten IDE drives to be attached to the DDA card. DDA disk drives may be configured through the EISA configuration utility to form "composite" drives consisting of several physical drives in a drive array. These composite drives may be one of three types of arrays in which data is striped across the individual drives. These are the standard array form, the mirrored array, and the guarded array, previously described

in connection with the first preferred embodiment. On each cable in the super cable assembly, there are two connectors, the inner connector and the outer connector at the end of the cable. The disk drives attached to the inner connectors form one band of drives and the ones connected to the outer connectors form another. There are at most five disk drives in one bank. The composite drives are then subject to one rule: a composite drive must consist of disk drives in a single bank. Mirroring changes the rule slightly. Each array in the mirrored pair must consist of disk drives in a single bank.

DDA performs many optimizations for disk I/O performance enhancement. These include:

Multiple I/O Threads DDA supports multiple outstanding I/Os on each logical drive, with operations on separate logical drives occurring concurrently.

Disk Cache The buffer manager implements a rudimentary store-thru cache.

This cache is fully associative and has a variable block size. Due to the small size of the current cache, lookups are very fast. There is a background process which attempts to control fragmentation of the cache. The cache can be turned off with the EISA configuration utility.

Readahead DDA keeps track of the addresses of the last n reads (where n is a programmable parameter). If a new read request comes in adjacent to any of the last n, then lookahead reads are buffered accordingly. The particular lookahead read strategy may be as simple as read an additional integral multiple of the requested read or more involved and depend upon the size of the read. For example, if a read request for sectors 10000–10005 follows four read requests after (n at least 4) a read request for sectors 9998–9999, then rather than just reading sectors 10000–10005, DDA will read sectors 10000–10017 (three times the requested size) in anitiepation of an imminent read request for sectors 10006–10017. Preferably n is set comparable to or greater than the maximum number of independent activities which may be underway. Thus if one thread is doing a sequential read the controller will perform readahead for optimization, but otherwise the controller will not normally do readahead.

Scheduling DDA currently performs no head scheduling. The request queues are strictly FIFO except for the read promotions discussed above.

Posted Writes If enabled, DDA returns a successful completion of a write as soon as it has filled a buffer with the write data from the host. This write is then queued onto the IDE drivers request queue, and the write occurs sometime in the future. This is basically an optimization for operating systems which do not support multi-threaded I/o.

Read and Write Combining Multiple reads to contiguous disk blocks are combined into one disk read, with the data being scattered/gathered to/from multiple buffers. This greatly reduces the disk latency and provides the benefits of scatter/gather operations without the associated protocol overhead. It also helps defray the performance degradation caused when the cache gets fragmented.

Read Promotion Read requests in the IDE request queue are promoted past write requests unless the read is for a block that the write is going to write to. Reads are not promoted past any other type of request.

Seek Swallowing Seeks in the IDE request queue are not performed if another request comes in behind them. This is safe because every operation has an implied seek. A sequence of N seeks will only execute the last one.

Firmware Patching

The DDA hardware provides 512 KB of ROM and 256 KB of RAM. The DDA firmware consists of approximately 256 KB of executable code and intialized data. Due to the relative imbalance of ROM and executable code size to RAM size, DDA implements an unusual approach to provide field downloadable firmware. The DDA firmware is divided into functional groups and each group further divided into successively smaller subgroups until each firmware function is sufficiently small in both scope and code size. References to these firmware functions is made exclusively through function pointers rather than direct calls. Using this technique, any firmware function may be entirely replaced without disrupting the main body of the firmware. Where a typical firmware initialization would consist of a single general initialization, DDA's firmware initializes in two distinct phases; first the public function pointers, known as patch pointers, are established, and then general initialization is performed. Using this technique, general initialization itself is fully patchable. Each functional group is responsible for implementing its own patch and initialization phases. A typical patch phase would look similar to the following:

```
emulPatch()
{
    emulISR = patch(DefISR);
    emulDecode = patch(DefDecode);
    emulReadAhead = patch(DefReadAhead);
    emulInit = patch(DefInit);
}
```

In this example, the Def* symbols represent function entry points found in the DDA ROM and the emul* symbols represent patch pointers. The patch( ) function returns a pointer to a function, either the pointer passed to it or a RAM-based replacement function. Code which desires to use any of the emul* functions in this example would call the function(s) through the patch pointers and would not know whether the actual service routine is running from ROM or from RAM.

Firmware patch routines may contain the following items: code (.text objects), constant data (data in the text segment), initialized data (.data objects), and uninitialized data (.bss objects). Objects in the .bss section will be initially zero-valued, and constant data may actually be changed at run-time. Patch routines may not call ROM routines directly, but must do so through patch pointers. If such a pointer does not exist, one can be created and placed in the .text or .data segments. Patch routines cannot know the addresses of objects in other patch routines. For this reason, a generic array of integers is reserved in the ROM data segment. This area can be used to allow several patch routines to share data.

At the beginning of the patch initialization phase, a firmware data structure is loaded from disk. This data structure is known as the patch record and is described below:

DDA ROM target version	4 bytes
patch version number	4 bytes
patch CRC	4 bytes
total length	4 bytes
patch block 1	variable
patch block 2	variable
patch block 3	variable
. . .	variable

The DDA ROM target version must match the actual ROM version. This field exists so that DDA can automatically

invalidate the patch record when its firmware ROMs are upgraded. The patch version number is a 32 bit integer providing a unique identifier for the patch record. This field is made available to host software after DDA initialization. The patch CRC is used by the firmware to verify the correctness of the patch record, and the patch length is used to calculate the actual patch CRC.

A patch block contains a everything necessary to replace a single ROM entry point. The structure of a patch block is defined as follows:

romAddr	4 bytes
loadAddr	4 bytes
nextPatch	4 bytes
codeSize	4 bytes
code	variable
fixup length	4 bytes
fixup list	variable

The romAddr field contains the address of the ROM routine the code body is designed to replace. The loadaddr field is reserved for use by the DDA firmware. The nextPatch field points to the next patch block, or contains zero if the patch block is the last one. The codeSize field contains the size of the code area in bytes. The code area contains the actual code and data that the patch block provides, and the fixup length and list provide the fixup information for the dynamic linker. The fixup list is an array of 32 bit integers containing the relative addresses within the code area that need address fixups. Only certain types of 960 intruction encoding may be used in a patch block, though all code generated by the firmware cross-compiler is supported.

When DDA powers up, it searchs the attached drives for a valid patch record. When it finds a valid record, it loads into low memory, chains all the patch blocks together, and zeros out the loadaddr fields of all the patch blocks. When the patcho function is called, the patch block chain is traversed. If a replacement function is found, it is copied into high memory, its fLxup addresses are resolved, and its new address is stored in loadaddr. If another call to patch( ) references the same replacement function, the loadaddr field will be non-zero and its value will simply be returned. This insures that, at most, only one copy of any replacement function is made. This is required because of the possibility of local data being associated with the replacement function.

DDA Fragmentation Mechanism

Defect Management, Request Fragmentation, and Queue Management

FIG. **2** shows ideal platters for disk drives **110–114**. In contrast, the physical disk drives may include platters with defective sectors, both defects arising from manufacture and defects "growing" after manufacturers testing and shipping, such as read/write head alignment drift. To compensate for manufacturing defects, the controller on each disk drive includes a lookup table for mapping defective sectors to spare sectors reserved for this purpose. That is, each disk drive appears to have its sectors logically identified by sequence (or by geometry) and this may correspond to the physical layout of the sectors provided there are no defective sectors; alternatively, there may be a simple geometrical mapping of logical to physical locations to accommodate disk drive standards such as 17 sectors per track. However, if a manufacturing defect appears in a sector, then this sector is remapped to a spare sector and the disk drive controller uses its lookup table of defects to treat logical access to this defective sector as physical access to such spare sector. Note that the error correction mechanism (e.g., CRC or ECC) for a sector will detect soft errors which may arise from a defect growing after manufacture, and the disk drive's controller will determine the correct sector data using the appropriate error correction algorithm and thus overcome such a defect. Such error correction and such remapping of defective sectors is transparent to the user (i.e., controller **100**) of the disk drive, although soft errors may be indicated. Conversely, if a hard error arises, as from a growing defect, then the disk drive's controller cannot recover the data and indicates an error to controller **100** for an attempted read or write.

Controller **100** modifies the simple remapping of bad sectors just described with the following remapping. If, for example, sector 121 develops a defect resulting in a hard error, then controller **100** computes the correct data for sector 121 using sectors 120 and 122–124 as previously described and then writes all five sectors' data to a new stripe of sectors 130–134; that is, each sector in the stripe is remapped to a spare sector on the same disk to form a new stripe. Controller **100** records both the remapping of logical to physical locations and the identity of the disk which developed the detected defect. Controller **100** then proceeds to service the original read/write request which uncovered the defect.

Description of Terms

In the following descriptions of fragment operations, snapshots of the request queue for each stage in the process will be provided. Each entry in the queue will be described by four fields plus a possible comment. Although a queue entry contains much more than four relevant fields, this representation is sufficient for our purposes. The first field is the request type. Request types will be described later. The second field is the command type. Command types will also be described later. The third field is the starting sector number. The sectors which compose the available space on a composite disk are addressed through a zero-based absolute sector number. A starting sector number of zero corresponds to the first available sector on a composite drive. The last field is the sector count. Each request which addresses composite disk data specifies the amount of data, in sectors, in this field.

Request Type

The first field of the request is the request type. The following is a description of the different request types that appear in this text. During the course of request processing, the DDA firmware may change a request's type.

ORIGINAL: A request that has been submitted to the disk driver by an outside source. This does not necessarily have to be the host; the DDA firmware submits requests to the disk driver for a variety of reasons.

FRAGMENT: A request that performs only a portion of the work necessary to complete an ORIGINAL request. A FRAGMENT is usually atomic.

FRAG-ORIG: When an ORIGINAL request is fragmented, its type is changed to FRAG-ORIG.

COMBINATION: A request consisting of multiple ORIGINAL requests that may be completable atomically. Sequential reads and writes may be combined to improve performance. When this occurs, a COMBINATION request is created and the ORIGINAL requests are converted into COMB-ORIG requests.

COMB-ORIG: When a COMBINATION request is created, the ORIGINAL requests which will be satisfied by the COMBINATION request are converted into COMB-ORIG requests.

FRAG-COMB: When a COMBINATION request cannot be satisfied atomically, it is fragmented and its type is changed to FRAG-COMB.

Command Type

The second field of the request is the command type. The following is a description of the different internal command types that appear in this text. It is not an exhaustive list of the internal commands supported in the DDA firmware. There are two classes of commands, atomic and META. An atomic command is one which can be completed in one step. META commands are commands which must be fragmented into atomic commands and are not executed themselves. META commands are not subject to remapping since they will be "cracked" into atomic commands which will be remapped if necessary. Note that atomic commands that are subject to remapping can still be fragmented as a result of the remap process. Therefore, "atomic" does not mean "non-fragmentable." Some commands are always atomic, others are always META, and still others are sometimes atomic and sometimes META

READ: A logical disk read operation. This request is atomic, it can be generated by the host and is subject to remapping.

WRITE: A logical disk write operation. In the case of a guarded composite drive, the command is implemented as a META command when necessary. Otherwise, it is atomic. This request can be generated by the host and is subject to remapping.

WRITEVERF: A logical disk write plus verify operation. This request performs a verify after write to help insure data integrity and is commonly submitted by host software. This request is implemented as a META command, since IDE drives do not generally provide WRITE plus VERIFY operations.

VERIFY: A logical disk verify operation. Essentially the same as a read, but the data is not actually transferred. It is used to test the integrity of disk sectors. This request is atomic, it can be generated by the host and is subject to remapping.

VERIFY-NR: Another logical disk verify operation. It differs from the verify operation in that the physical disks are instructed to disable retries. It is used to help isolate errors. This request is atomic, it cannot be generated by the host and it is subject to remapping.

RESET: A logical disk reset operation. This request may be submitted by the host in the event of an error and is implemented as a META command.

RAWRESET: An atomic internal command used by the disk driver to help perform a disk reset.

SETMULT: An atomic internal command used by the disk driver to help perform a disk reset.

SETPARMS: An atomic internal command used by the disk driver to help perform a disk reset.

SETBUFF: An atomic internal command used by the disk driver to help perform a disk reset.

MAPTHIS: This command performs a manual remap operation. This request is only generated by the host and is implemented as a META command.

RMW-READ: An atomic internal command used by the disk driver to implement guarded writes. It is similar to a normal read operation except that it associated with guarded write commands. This request cannot be generated by the host but is subject to remapping.

RMW-WRITE: An atomic internal command used by the disk driver to implement guarded writes. It is similar to a normal write operation except that it associated with guarded write commands and must know how to generate parity drive data using data previously generated by a RMW-READ command. This request cannot be generated by the host but is subject to remapping.

WAP-READ: An atomic internal command used by the disk driver to implement sector remapping. It is similar to a normal read operation except that it is not subject to remapping and it must be able to recover lost sectors. This request cannot be generated by the host.

MAP-WRITE: An atomic internal command used by the disk driver to implement sector remapping. It is similar to a normal write operation except that it is not subject to remapping and the request is always stripe aligned. This request cannot be generated by the host.

READ-ALL: An atomic internal command used by the disk driver to retrieve controller information. This command is not subject to remapping.

READ-SAME: An atomic internal command used by the disk driver to retrieve controller information. This command is not subject to remapping.

WRITE-SAME: An atomic internal command used by the disk driver to store controller information. This command is not subject to remapping.

WRITE-RMAP: An atomic internal command used by the disk driver to implement the final phase of a remap or uniap operation. It stores the remap table information on the disk. This command is not subject to remapping.

WRITE-DFL: An atomic internal command used by the disk driver to store physical disk defect lists. This request cannot be generated by the host and is not subject to remapping.

WRITE-STAT: An atomic internal command used by the disk driver to store composite drive status information. This command is not subject to remapping.

Fundamental Operations

The following is a list of operations that the DDA disk driver performs 5 using the fragmentation mechanism. It is not an exhaustive list, though it encompasses all of the basic uses of fragmentation. An example of a real request that uses many of these operations can be provided but it would be very long-winded.

Error Isolation

When an unrecoverable error occurs on a multi-sector operation, DDA uses the fragmentation mechanism to isolate the bad sector(s). In the following example, the composite drive is a two drive array and the bad sector is sector 2.

Initial Request

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	0	8	Head

The host has issued a read request. The I)DA firmware attempts the read but receives an error from one of the

drives. The DDA firmware attempts to isolate the failure using a VERIFY-NR command.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	VERIFY-NR	0	8	Head
FRAG-ORIG	READ	0	8	

The verify operation reports that sector 2 fails. Since a remap is not possible, we will reset the drives and retry the verify. If we were able to remap, we would dequeue the verify, push the remap operation which will be described later, and the push the reset operaetions.

After Verify Failure

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RAWRESET	—	—	Part of Reset - Head
FRAGMENT	SETPARMS	—	—	Part of Reset
FRAGMENT	SETBUFF	—	—	Part of Reset
FRAGMENT	SETMULT	—	—	Part of Reset
FRAGMENT	VERIFY-NR	0	8	
FRAG-ORIG	READ	0	8	

DDA then executes the four requests which implement a drive reset.

The queue looks the same as it did after the initial fragmentation, but a retry counter, not shown, prevents us from entering an endless loop.

After Reset

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	VERIFY-NR	0	8	Head
FRAG-ORIG	READ	0	8	

The verify operation again reports that sector 2 fails. DDA calls the completion function of the ORIGINAL read and reports the error. The queue is now empty.

Sector Remapping

Sector remapping involves two operations: recovering data and 10 transferring it to a newly allocated location (dynamic sector remapping), and converting all subsequent references to the remapped location (runtime remapping).

Dynamic Sector Remapping

In the following example, we have a multi-sector read request that references a sector that has gone bad. The composite drive is a two-plus-parity guarded array and the bad sector is sector 2. This example is identical to the example in the Error Isolation section except that DDA can recover the lost data since the composite drive contains redundancy.

Initial Request

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	0	8	Head

The host has issued a read request. The DDA firmware attempts the read but receives an error one of the drives. The DDA firmware attempts to isolate the failure using a VERIFY-NR command.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	VERIFY-NR	0	8	Head
FRAG-ORIG	READ	0	8	

The verify operation reports that sector 2 fails. DDA decides to remap the stripe containing sector 2 to a reserved area at sector 10000.

After Isolation of Read Error

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RAWRESET	—	—	Part of Reset - Head
FRAGMENT	SETPARMS	—	—	Part of Reset
FRAGMENT	SETBUFF	—	—	Part of Reset
FRAGMENT	SETMULT	—	—	Part of Reset
FRAGMENT	MAP-READ	2	2	
FRAGMENT	MAP-WRITE	10000	2	
FRAGMENT	WRITE-RMAP	—	—	Write Remap Table
FRAGMENT	WRITE-DFL	—	—	Write Defect Lists
FRAG-ORIG	READ	0	8	

DDA then executes the four requests which implement a drive reset.

After Isolation of Read Error

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	MAP-READ	2	2	Head
FRAGMENT	MAP-WRITE	10000	2	
FRAGMENT	WRITE-RMAP	—	—	Write Remap Table
FRAGMENT	WRITE-DFL	—	—	Write Defect Lists
FRAG-ORIG	READ	0	8	

The firmware launches the MAP-READ command knowing that sector 2 is bad and must be recovered. The rebuilt data is placed in private buffers that the MAP-WRITE command will use to write the recovered data to the remap location.

After Recovery of Lost Data

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	MAP-WRITE	10000	2	Head
FRAGMENT	WRITE-RMAP	—	—	Write Remap Table
FRAGMENT	WRITE-DFL	—	—	Write Defect Lists
FRAG-ORIG	READ	0	8	

The firmware launches the MAP-WRITE command to write the recovered data to the remap location.

After Write of Recovered Data to Remap Area

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-RMAP	—	—	Head
FRAGMENT	WRITE-DFL	—	—	Write Defect Lists
FRAG-ORIG	READ	0	8	

The firmware now writes the new remap table onto a reserved area of the disks to complete the logical remap operation.

After Write of Remap Table

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Head
FRAG-ORIG	READ	0	8	

The firmware now updates the physical defect lists of all physical drives containing defective sectors. In this example, only one drive contained an error. If more than one drive contained an error, the WRITE-DFL command itself would be fragmented so that each WRITE-DFL command would update only one physical drive. See the "Write Physical Defects" section for further details.

After Write of Physical Defect Lists

Req Type	Command	Sec #	Cnt	Comments
FRAG-ORIG	READ	0	8	Head

Now that the remap is complete, the queue will be restored to its original state and we will start over. This will allow the read to complete using the Runtime Remapping method described below.

After Remap Complete

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	0	8	Head

Manual Sector Remapping

In the following example, the host has requested that a sector remap be performed. Typically, this is done because a verify request reported an error. Without considering how the data is recovered, we will follow the manual remap process. The composite drive is a two drive non-redundant array and the sector to be remapped is sector 2.

Initial Request

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	MAPTHIS	2	1	Head

DDA decides to remap the stripe containing sectors 2 and 3 to a reserved area at sector 10000. Although not shown here, DDA remembers that the first drive in the data stripe is responsible for the remap since that is what the host specified. This is important to the defect list management and to the data recovery routines. Data on sector 3 must be recovered since is not involved in the original host request.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	MAP-READ	2	2	Head
FRAGMENT	MAP-WRITE	10000	2	
FRAGMENT	WRITE-RMAP	—	—	Write Remap Table
FRAGMENT	WRITE-DFL	—	—	Drive 1 Defect List
FRAG-ORIG	MAPTHIS	2	1	

The firmware launches the MAP-READ command knowing that sector 1 is bad and does not have to be recovered. It will recover sector 1 if at all possible, however. The rebuilt data is placed in private buffers that the MAP-WRITE command will use to write the recovered data to the remap location.

After Recovery of Data

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	MAP-WRITE	10000	2	Head
FRAGMENT	WRITE-RMAP	—	—	Write Remap Table
FRAGMENT	WRITE-DFL	—	—	Drive 1 Defect List
FRAG-ORIG	MAPTHIS	2	1	

The firmware launches the MAP-WRITE command to write the recovered data to the remap location.

After Write of Recovered Data to Remap Area

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-RMAP	—	—	Head
FRAGMENT	WRITE-DFL	—	—	Drive 1 Defect List
FRAG-ORIG	MAPTHIS	2	1	

The firmware now writes the new remap table onto a reserved area of the disks to complete the logical remap operation.

After Write of Remap Table

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Head
FRAG-ORIG	MAPTHIS	2	1	

The firmware now updates the physical defect list on drivel and calls the completion function for the MAPTHIS command. The queue is now empty.

Runtime Remapping

In the following example, we have a multi-sector read request that references a remap location. The composite drive is a two-plus-parity guarded array and the remapped sectors are sectors 2 through 3 which are remapped to sectors 10000 through 10001. This example describes the situation which would exist after the remapping describe above.

Initial Request

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	0	8	Head

The host has issued a read request. The DDA firmware recognizes the remapped area and fragments the request.

After First Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	READ	0	2	Head
FRAGMENT	READ	2	6	Remainder
FRAG-ORIG	READ	0	8	

The DDA firmware performs the first read request since it is now contiguous.

After Read of First Contiguous Block

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	READ	2	6	Head
FRAG-ORIG	READ	0	8	

The DDA firmware again recognizes the remapped area and fragments the request again.

After Second Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	READ	10000	2	Head
FRAGMENT	READ	4	4	Remainder
FRAG-ORIG	READ	0	8	

The DDA firmware performs the read request on the remapped area since it is now contiguous.

After Read of Second Contiguous Block

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	READ	4	4	Head
FRAG-ORIG	READ	0	8	

The DDA firmware performs the read request on the remaining area since it is also contiguous.

After Read of Last Contiguous Block

Req Type	Command	Sec #	Cnt	Comments
FRAG-ORIG	READ	0	8	Head

The DDA firmware recognizes the FRAG-ORIG type at the head of the queue which signifies the completion of the ORIGINAL host request. It calls the completion function associated with the ORIGINAL request.

Write Physical Defects

When a remap occurs, DDA updates the physical defect lists of those drives whose errors caused the remap to happen. Since multiple drive errors are unlikely but could occur, DDA reserves just enough memory to process one drive at a time and uses the fragmentation process to support multiple drive errors. In the following example, three drives, drives **1, 2,** and **3,** contained errors which led to the remap. Since the WRITE-DFL can only exist as a fragment, that is all that is shown in this example.

After Write of Remap Table

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Drives 1, 2 and 3
FRAG-ORIG	READ	0	8	
ORIGINAL	READ	0	8	Failing Command

We recognize that multiple drives must be updated by this request, so we fragment it.

After Fragmentation 1

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Drive 1
FRAGMENT	WRITE-DFL	—	—	Drives 2 and 3
FRAG-ORIG	READ	0	8	
ORIGINAL	READ	0	8	Failing Command

We write the defect list to drive **1** since only one drive is specified by the first request.

After Write 1

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Drives 2 and 3
FRAG-ORIG	READ	0	8	
ORIGINAL	READ	0	8	Failing Command

We recognize that multiple drives must be updated by this request, so we fragment it.

After Fragmentation 2

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Drive 2
FRAGMENT	WRITE-DFL	—	—	Drive 3
FRAG-ORIG	READ	0	8	
ORIGINAL	READ	0	8	Failing Command

We write the defect list to drive **2** since only one drive is specified by the first request.

After Write 2

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-DFL	—	—	Drive 3
FRAG-ORIG	READ	0	8	
ORIGINAL	READ	0	8	Failing Command

We write the defect list to drive **3** since only one drive is specified by the first request.

After Write 3

Req Type	Command	Sec #	Cnt	Comments
FRAG-ORIG	READ	0	8	Head
ORIGINAL	READ	0	8	Failing Command

We have completed the WRITE-DFL operation.

Read/Write Combining

When multiple sequential read or write requests exist on the drive queue, DDA will combine those requests into single, large requests to enhance performance. This is especially effective for small writes on guarded arrays. If an error occurs, the requests are decombined and run as ORIGINALs to simplify error handling. In the following example, multiple, sequential disk reads are enqueued. The composite drive type is unimportant, but the maximum transfer size of the composite disk is 128.

Initial Queue State

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	0	4	Head
ORIGINAL	READ	4	4	
ORIGINAL	READ	8	4	
ORIGINAL	READ	12	128	

DDA recognizes that the first four requests on the queue are sequential reads. It combines the first three, but not the fourth request since the transfer size would become greater than the maximum allowable size.

After Combination of Requests

Req Type	Command	Sec #	Cnt	Comments
COMBINATION	READ	0	12	Head
COMB-ORIG	READ	0	4	
COMB-ORIG	READ	4	4	
COMB-ORIG	READ	8	4	
ORIGINAL	READ	12	128	Size too Large

DDA performs the combination read successfully. It scans the queue. and calls the completion function for every request satisfied by the combined read.

After Completion of Combined Read

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	READ	12	128	Head

Guarded Write Processing

Writes to guarded composite drives are complicated by the fact that parity sectors must be generated for each data stripe. When write encompass full data stripes, the parity sectors for those stripes can be generated directly. In other cases, a Read-Modify-Write (RMW) operation must be performed. RMW operations are performed through fragmentation. Therefore, the WRITE command itself is conditionally a META command. In the following example, several writes are enqueued on a two-plus-parity guarded composite drive. The maximum transfer size of the drive is 128.

Initial Queue State

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	0	2	Head
ORIGINAL	WRITE	4	1	
ORIGINAL	WRITE	8	3	
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

The first request cannot be combined but is stripe aligned. DDA fragments it into a single RMW-WRITE operation to simplify error handling.

After Fragment 1

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-WRITE	0	2	Head
FRAG-ORIG	WRITE	0	2	Was ORIGINAL
ORIGINAL	WRITE	4	1	
ORIGINAL	WRITE	8	3	
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

DDA the executes the RMW-WRITE operation and calls the WRITE completion function.

After Write 1

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	4	1	Head
ORIGINAL	WRITE	8	3	
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

The next request cannot be combined and is not stripe aligned. DDA fragments it into read and write phases. The write phase is capable of generating the parity data on demand. The read phase must start and end on a stripe boundary and must entirely contain the data to be written.

After Fragment 2

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-READ	4	2	Head
FRAGMENT	RMW-WRITE	4	1	
FRAG-ORIG	WRITE	4	1	Was ORIGINAL
ORIGINAL	WRITE	8	3	
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

DDA executes the RMW-READ operation.

After Read 2

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-WRITE	4	1	Head
FRAG-ORIG	WRITE	4	1	Was ORIGINAL
ORIGINAL	WRITE	8	3	

-continued

After Read 2

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

DDA the executes the RMW-WRITE operation and calls the WRITE completion function.

After Write 2

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	8	3	Head
ORIGINAL	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

DDA Combines the next two writes.

After Combine 3

Req Type	Command	Sec #	Cnt	Comments
COMBINATION	WRITE	8	8	Head
COMB-ORIG	WRITE	8	3	
COMB-ORIG	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

The combined write is stripe-aligned so it is fragmented into a RMW-WRITE operation.

After Fragment 3

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-WRITE	8	8	Head
FRAG-COMB	WRITE	8	8	Was COMBINATION
COMB-ORIG	WRITE	8	3	
COMB-ORIG	WRITE	11	5	
ORIGINAL	WRITE	21	2	
ORIGINAL	WRITE	23	4	

DDA the executes the RMW-WRITE operation, recognizes the completion of a COMBINATION, and calls the completion functions that are satisfied by the combined write.

After Write 3

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	21	2	Head
ORIGINAL	WRITE	23	4	

DDA Combines the next two writes.

After Combine 4

Req Type	Command	Sec #	Cnt	Comments
COMBINATION	WRITE	21	6	Head
COMB-ORIG	WRITE	21	2	
COMB-ORIG	WRITE	23	4	

The combined write is not stripe-aligned so it is fragmented into read and write phases. The read phase must start and end on a stripe boundry and must entirely contain the data to be written.

After Fragment 4

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-READ	20	8	Head
FRAGMENT	RMW-WRITE	21	6	
FRAG-COMB	WRITE	21	6	
COMB-ORIG	WRITE	21	2	
COMB-ORIG	WRITE	23	4	

DDA executes the RMW-READ operation

After Read 4

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RMW-WRITE	21	6	Head
FRAG-COMB	WRITE	21	6	
COMB-ORIG	WRITE	21	2	
COMB-ORIG	WRITE	23	4	

DDA then executes the RMW-WRITE operation, recognizes the completion of a COMBINATION, and calls the completion functions that are satisfied by the combined write. The queue is now empty.

Write+Verify Operation

Host software frequently desires to perform a verify operation after each write operation to insure that the data written is readable. This request appears as a special WRITEVERF command on the DDA drive queue. Since IDE drives do not support verify-after-write operations, DDA uses the fragmentation operation to perform the WRITEVERF command. Therefore, WRITEVERF is always a META command and takes advantage of the fact that WRITE and VERIFY are already implemented to provide this function in essentially no additional code.

Initial Queue State

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITEVERF	0	4	Head

DDA fragments the WRITEVERF command into WRITE and VERIFY commands.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE	0	4	Head
FRAGMENT	VERIFY	0	4	
FRAG-ORIG	WRITEVERF	0	4	

DDA performs the WRITE.

After Write

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	VERIFY	0	4	Head
FRAG-ORIG	WRITEVERF	0	4	

DDA performs the VERIFY and calls the completion function. The queue is now empty.

Host Reset Processing

Host software frequently sends a reset command to the controller to set the controller into a known state either on power-up or following a drive error. Resets involve a number of consecutive operations. Therefore, the RESET command is implemented as a META command. DDA does not generate RESET commands internally. The following is an example of a RESET command.

Initial Queue State

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	RESET	—	—	Head

DDA fragments the RESET command into the required pieces.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	RAWRESET	—	—	Head
FRAGMENT	SETPARMS	—	—	
FRAGMENT	SETBUFF	—	—	
FRAGMENT	SETMULT	—	—	
FRAG-ORIG	RESET	—	—	

Each fragment on the queue at this point is atomic. The commands on the queue are the minimum required by the DDA firmware to reset the physical drives as set them into the correct state.

Drive Status Update

In a redundant composite drive, physical drives can fail without causing the composite drive to become unusable. When this occurs, physical drive status must be tracked by DDA. In the event of a total drive failure, DDA continues to allow data requests on the drive until a write involving a failed disk is requested. At this point, DDA must mark the disk as permanently unusable before performing the write since the data on the failed disk will become invalid. DDA stores this information on the remaining physical disk which make up the composite drive.

In the following example, the first drive of a two-plus-parity guarded array has failed. The firmware continues to operate until a write operation involving drive **1** appears at the head of the queue.

Initial Queue State

Req Type	Command	Sec #	Cnt	Comments
ORIGINAL	WRITE	0	4	Head

DDA fragments the WRITE command into a WRITE and a WRITE-STAT command.

After Fragmentation

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE-STAT	0	4	Head
FRAGMENT	WRITE	0	4	Copy of ORIGINAL
FRAG-ORIG	WRITE	0	4	

DDA performs the WRITE-STAT command to invalidate drive **1**. Notice that the actual host WRITE cannot occur until after the successful completion of the WRITE-STAT command. This insures that an ill-timed power failure cannot corrupt the state of the host's data.

After Status is Updated

Req Type	Command	Sec #	Cnt	Comments
FRAGMENT	WRITE	0	4	Head
FRAG-ORIG	WRITE	0	4	

DDA performs the WRITE that was originally requested by executing the FRAGMENT at the head of the queue. It then calls the completion function for the ORIGINAL write. The queue is now empty.

On every array access, controller **100** searches its remap tables to determine whether the requested access will include one or more of the remapped stripes. If a remapped stripe occurs in a requested access, then the requested access is split into multiple accesses in which the accesses to relocated sectors are remapped to their new locations. For example, a read or write requesting access to the stripe of sectors **120–124** plus the stripes of the sectors immediately preceding and following each of sectors **120–124** on the same track on each of the platters would be split from a simple operation accessing three successive stripes to one accessing the remapped sectors **120–124** (e.g., sectors **140–144**) plus one accessing the original immediately preceding and following sectors. Thus the read/write heads would seek two different tracks during the operation. The following example more explicitly shows the queue in controller **100** during a multisector read that includes a sector with a grown defect; the first portion of the example shows the dynamic remapping.

Presume that the host requests a read of 20 sectors beginning at sector 0^2 with array setup for redundancy; thus five stripes of data are to be read. Also, presume that sector

9 has developed a hard error defect. Thus the initial request in the queue looks like (in simplified version for clarity):

[2](For simplicity, logical sector numberings will be used in this example.)

Request Type	Command	Sector No.	Sector count
Original	Read	0	20

The first column shows that the Read from the host was an Original request in the sense that it derived from an outside source. The second column illustrates the command from controller **100** to the disk drives **110–114**. The third column states the starting sector number (as seen by the host), and the fourth column indicates the number of contiguous sectors to be read. Now bad sector **9** cannot be read by its disk drive (disk drive **111**), so controller **100** receives an error message from this disk drive. Controller **100** then attempts to isolate the error by sending a Verify-NR command to the disk drives **110–114** which is a logical disk drive verify operation to test integrity of sectors; the queue with this Verify-NR command now at the head of the queue appears as:

Request Type	Command	Sector No.	Sector count
Fragment	Verify-NR	0	20
Frag-Orig	Read	0	20

The Verify-NR command is called a Fragment type request in that the original request has been fragmented in order to isolate the error; the change of type from Original to Frag-Orig for the host's Read request reflects this fragmentation. The Verify-NR operation reports to controller **100** that sector **9** has failed. Controller **100** then selects a reserved stripe, say sectors **50,000** through **50,003** plus the corresponding unnumbered parity sector, for remapping the stripe with failed sector **9**; that is, the stripe formed by sectors **8** though **11** plus parity sector. Controller **100** then inserts the commands for the remapping into queue which then appears as:

Request Type	Command	Sector No.	Sector count
Fragment	Rawreset	—	—
Fragment	Setparms	—	—
Fragment	Setbuff	—	—
Fragment	Setmult	—	—
Fragment	Map-Read	8	4
Fragment	Map-Write	50000	4
Fragment	Wrt-Rmp	—	—
Fragment	Wrt-DFL	—	—
Frag-Orig	Read	0	20

The first four commands inserted into the queue provide a reset of the disk drives; and after the disk drives execute these commands the queue looks like:

Request Type	Command	Sector No.	Sector count
Fragment	Map-Read	8	4
Fragment	Map-Write	50000	4
Fragment	Wrt-Rmp	—	—
Fragment	Wrt-DFL	—	—
Frag-Orig	Read	0	20

Controller **100** knows that sector **9** has failed and the Map-Read command for the stripe containing sector **9** and uses the parity sector to rebuild the data lost in sector **9** and stores the entire stripe of data in buffers. Thus the queue is now:

Request Type	Command	Sector No.	Sector count
Fragment	Map-Write	50000	4
Fragment	Wrt-Rmp	—	—
Fragment	Wrt-DFL	—	—
Frag-Orig	Read	0	20

Controller **100** now launches the Map-Write command to write the rebuilt data in the buffers to the stripe starting at sector **50,000**, so the queue becomes:

Request Type	Command	Sector No.	Sector count
Fragment	Wrt-Rmp	—	—
Fragment	Wrt-DFL	—	—
Frag-Orig	Read	0	20

Controller **100** uses Writ-Rmp to write the new remap table onto a reserved area of the disk drives to complete the logical remap operation. Thus the queue is:

Request Type	Command	Sector No.	Sector count
Fragment	Wrt-DFL	—	—
Frag-Orig	Read	0	20

The Wrt-DFL command causes controller **100** to update its physical defect lists of all physical drives containing defective sectors. In this example, only one disk drive contained an error; if more than one drive contained an error, then the Wrt-DFL command itself would be fragmented so that each Wrt-DFL command would update only one physical disk drive error. Now the queue contains a single command:

Request Type	Command	Sector No.	Sector count
Frag-Orig	Read	0	20

Thus the remapping is complete and now the queue will have been restored to its original state and the host's Read request will be recast as an Original request. The execution of the Read request will now use runtime remapping because controller **100** will see that one of the stripes to be read has been remapped. In particular, controller **100** will again fragment the host's Read request and insert commands into the queue as follows:

Request Type	Command	Sector No.	Sector count
Fragment	Read	0	8
Fragment	Read	8	12
Frag-Orig	Read	0	20

The first Read of sectors **0** through **7** can proceed because the stripes are contiguous, and the queue becomes:

Request Type	Command	Sector No.	Sector count
Fragment	Read	8	12
Frag-Orig	Read	0	20

The remapped stripe causes a further fragmentation:

Request Type	Command	Sector No.	Sector count
Fragment	Read	50000	4
Fragment	Read	12	8
Frag-Orig	Read	0	20

The disk drives execute the Read of the remapped stripe to leave the queue as:

Request Type	Command	Sector No.	Sector count
Fragment	Read	12	8
Frag-Orig	Read	0	20

The Read again executes because the stripes are now contiguous. Thus the queue is:

Request Type	Command	Sector No.	Sector count
Frag-Orig	Read	0	20

Controller **100** recognizes the Frag-Orig at the head of the queue signifies completion of the original Read request from the host and calls the completion function associated with the original Read request.

Controller **100** can also associate multiple operations with a single request from the host in order to decompose a complex request to simple, fully restartable requests sequences for error handling.

Controller **100** will combine multiple sequential read or write requests in the controller **100** queue in order to access increase efficiency by having larger blocks read or written. For example, if the maximum transfer size of controller **100** is 128 sectors (64 KB) due to the size of its buffers, and if the following requests are pending in the request queue:

Request Type	Command	Sector No.	Sector count
Original	Read	0	11
Original	Read	11	17
Original	Read	28	2
Original	Read	30	110
Original	Read	140	10

Controller **100** recognizes the five requests as sequential, but cannot combine them all because this would exceed the transfer size. Thus controller **100** combines the first three requests and the queue appears as:

Request Type	Command	Sector No.	Sector count
Combination	Read	0	30
Comb-Orig	Read	0	11
Comb-Orig	Read	11	17
Comb-Orig	Read	28	2

-continued

Request Type	Command	Sector No.	Sector count
Original	Read	30	110
Original	Read	140	10

Controller **100** successfully performs the combination Read, scans the queue, and calls the completion function for every request satisfied by the combination Read. Then controller **100** continues with the queue looking like

Request Type	Command	Sector No.	Sector count
Original	Read	30	110
Original	Read	140	10

Controller **100** combines these two reads as before and reads them together.

Scatter/Scatter

The system of the presently preferred embodiment provides "scatter/scatter" accesses, in which both the physical locations of data in host memory and the physical locations of data on the disks can be discontiguous. That is, the host can send a single request to launch such a scatter/scatter transfer. Arguments to such a transfer request would include: a pointer to a list of transfer counts and addresses in host memory containing the data to be transferred; the length of that list; and the starting logical address on the disk for transfer.

Note that the host need not know the configuration that the data array will have on the disk.

Skipped blocks in a scatter-scatter request are specified by a data address value of −1. Thus, when a block must be skipped, the controller enqueues a "nop" (no-operation) request. Note that the presently preferred embodiment enqueues these nop requests, if needed, even if the data transferred is in contiguous addresses on the host memory side.

Any disk operation, in the presently preferred embodiment, is limited to a set maximum number of blocks of logical disk address space (currently 256). Thus, no scatter/scatter request can cover more than 254 skipped blocks.

The scatter-handling operations just described are implemented, in the presently preferred embodiment, using the controller's native mode described below.

Controller **100**, via bus master **108**, may also write the data obtained from such combined requests to physically scattered portions of the host RAM. That is, a scatter/scatter as suggested in FIG. **12** with the host requesting reads into three separated portions of its RAM and for data on three stripes with noncontiguous sectors but with sectors on a single track in each disk drive. The read requests are combined by controller **100** because the sectors can all be read in a single revolution of the disk drives' platters, and the transfer back to the RAM can be accomplished by bus master **108**. Note that the usual scatter/gather consists of gathering scattered portions of host RAM to contiguous sectors on disk, or gathers scattered sectors on disk to contiguous portions of host RAM.

FIG. **13** illustrates a simple scatter/scatter read in a format analogous to that of FIG. **8**. In particular, the example in FIG. **13** starts with a request from the host to read from disk **0** the data in logical sectors **100–101** (1 KB) and put it into 1 K of host memory beginning at address 2000 (in

hexadecimal) plus the data in logical sectors **104–105** (1KB) into 1 K of host memory locations beginning at address 3000. The request has the format of an address (1000) for a pointer which points to a list of the addresses 2000 and 3000 together with the count of sectors for each; the list also indicates the skipping of sectors **102–103** by the address −1 and count of 2 as the second list entry. Controller **100** decodes this request and allocates six read buffers (one for each of sectors **100–105**) plus a disk request (dskRequest) and three transfer requests (xferRequest); note that a fork counter is set to three and a join counter to zero to keep track of the completions of the three transfer requests. The disk request has a READ command (disk to buffer) for the read of six contiguous sectors beginning at sector **100** into 3 K of contiguous controller **100** buffer memory starting at address 1001000 (hexadecimal). Thus this is a single read which includes the desired data in sectors **100–101** and **104–105** plus the not needed data in sectors **102–103**. The data from sectors **100–101** goes into buffer memory at addresses 1001000–10013FF, the data from sectors **102–103** goes into buffer memory at addresses 1001400–10017FF, and the data from sectors **104–105** goes into buffer memory at addresses 1001800–1001BFF. There is one of the three transfer requests for each of these three address ranges. The transfer requests for the desired data each has a READ command (buffer to host) with the target host memory starting address, the buffer starting address, and the size count (in terms of sectors); that is, the first transfer request has beginning target address 2000 and buffer starting address 1001000, and the third transfer request has beginning target address 300 and buffer starting address 1001800. The transfer request for the unneeded data has a NOP (no operation) command instead of the READ command and has no target host memory address, but does have the starting buffer address 1001400. As each of the transfer requests completes, the corresponding buffer memory becomes free and available for use by other processes and the join count increments.

Note that controller **100** extends the scattered read from the logical disk into a single large read of contiguous sectors into buffers and then suppressing the unwanted data without changing the request structures by simply inserting a NOP command in place of the READ command for the buffer to host transfer.

Guarded Optimization

Writes in guarded configuration are actually read-modify-write (RMW) operations. Note that RMW operations are an example of the use of "fence pointers" as above. As noted, when doing a small write in guarded configuration, the DDA accesses either of two complementary sets of drives, depending on which is coming up sooner.

Rebuilding

In rebuilding data, a problem is that there is likely to be a large seek time going over to the rebuild data area. Thus, it is desirable to avoid thrashing going to and from the rebuild data area.

The DDA has two relevant tunable parameters (see FIG. 6):

1) Amount of idle time before we initiate a rebuild;

2) maximum number of requests which may be processed before a rebuild operation is forced to occur.

The advantage of this is that rebuild is guaranteed to complete within some determined time.

Rebuild operations too are implemented using the above queue-managment architectures. Fence markers can be used to protect a minimum-length rebuild operation from interruption by new requests.

Relevance to Error-handling and Recovery

When recovering from an error, a failed drive must be reset. Fragmentation queue-handling is used for this too: a reset operation request is pushed onto the stack.

It may happen that an error occurs during recovery from an error. The advanatage of fragmentation into atomic operations is that the DDA can cope with this without having to create nested error handling routines (which are a nuisance).

Hardware Architecture

Following is additional detail on the actual hardware implementation and register assignments used.

80960KA Side Global Memory I/O Map

Following is the present assignment of this map:

Address	Function	Notes
0x00000000 - 0x000FFFFF	ROM	not all populated
0x01000000 - 0x0100FFFF	SRAM	Normal Addressing
0x01040000 - 0x0104FFFF	SRAM	Permuted High
0x01080000 - 0x0108FFFF	SRAM	Permuted Low
0x03000000 - 0x03FFFFFF	Drives	Special mapping (see below)
0x04000000 - 0x04FFFFFF	Local I/O	See below
0x08000000 - 0x08FFFFFF	DRAM	Normal Addressing
0x09000000 - 0X09FFFFFF	DRAM	Permuted High
0x0A000000 - 0x0AFFFFFF	DRAM	Permuted Low
0xFF000000 - 0xFFFFFFFF	Intel Reserved	IACs, etc.

The address spaces are incompletely decoded, and a register or memory location may be aliased to many other places.

ROM map

Static RAM map

The DDA contains 256K bytes of static RAM. It is made out of eight 32KX8 SRAMs which are interleaved to achieve one wait state bursts on the 80960 LBUS and on the BMICs TDAT bus.

Dynamic RAM Map

The dynamic RAM map reserves three 16 Mbyte regions for the optional cache which behave like the static RAM regions described above.

Dual Port RAM Map

The register emulation works by dual porting the first 32 bytes of the static RAM mentioned above, although all 32 bytes are not defined, the undefined bytes in this region are reserved and will get trashed by the hardware.

AHA-1542 Mode:

EISA Address [1 . . . 0]	RAM Address Read	RAM Address Write
0x0	0x0100002	0x01000004
0x1	0x0100006	0x01000008
0x2	0x010000A	0x0100000C
0x3	0X010000E	0x01000010

After a write of 0x80, 0x40, or 0x20 to EISA address [1 . . . 2]==0 (i.e. DDA address (0x0l000004), EISA reads from EISA address [1 . . . 2]==2 (nominally DDA address 0x010000A) will come from DDA address 0x0100001C until RE2 is strobed low. This allows AHA resets to automatically clear the interrupt status.

IDE Drive Map

The mapping of the IDE drive task files has been optimized so that the 80960 can perform load and store multiples

in which all the drives can be accessed in a single instruction which results in a burst access on the 80960's LBUS. This burst access will allow the hardware to interleave the accesses to the IDE drives and get a 1 clock access on the second two drives compared to a three clock access for the first two drives. The drives are accessed two drives per word on typical operations. All drives are also aliased to another location in which all their registers overlap. This is write only and is used to send the same command to up to five drives in one access.

The memory map for the drives is contained in a 256 byte region which is depicted in the following table.

	Drive 3	Drive 2	Drive 1	Drive 0	Drive 4	All Drives
A3–A0 A7–A4	0b0110	0b0100	0b0010	0b0000	0b1010	0b11x0
0b0000	Data	Data	Data	Data	Data	Data
0b0001	Error	Error	Error	Error	Error	Error
0b0010	Sec Cnt	Sec Cnt	Sec Cnt	Sec Cnt	Sec Cnt	Sec Cnt
0b0011	Sec Num	Sec Num	Sec Num	Sec Num	Sec Num	Sec Num
0b0100	Cyl Low	Cyl Low	Cyl Low	Cyl Low	Cyl Low	Cyl Low
0b0101	Cyl High	Cyl High	Cyl High	Cyl High	Cyl High	Cyl High
0b0110	Drive	Drive	Drive	Drive	Drive	Drive
0b0111	Status	Status	Status	Status	Status	Status
0b1110	Alt Stat	Alt Stat	Alt Stat	Alt Stat	Alt Stat	Alt Stat
0b1111	Drv Addr	Drv Addr	Drv Addr	Drv Addr	Drv Addr	Drv Addr

Drives **0** and **2** are attached to the low 16 bits of the data bus, drives **1**, **3**, and **4** are connected to the high 16 bits of the data bus. When writing to all drives, the write data must be duplicated on the high and low halves of the data bus by software.

There are two drives on each cable. Drives **0** through **4** are accessed on cables **1** through **5** correspondingly, and drives **5** through **9** are also accessed on cables **1** through **5** correspondingly. The upper five drives are addressed at 0x00900 plus the addresses in the above table.

BMIC Map

The BMIC occupies a block of 16 bytes in the address space.

Address	Register	Size	Function
0x04000000	LDR	8 bits	Local Data Register
0x04000004	UR	8 bits	Local Index Register
0x04000008	LSR	8 bits	Local Status/Control Register
0x0400000C		8 bits	Intel Reserved

The BMIC uses an indexed I/O. For individual register index locations see the Intel documentation.

Address	Register	Size	Function
0x04000101		8 bits	
0x04000105		8 bits	
0x04000109		8 bits	
0x0400010D		8 bits	

DDA local I/O Ports

The DDA local I/O ports are located at:

Address	Register	Size	Function
0x04000202	INT-STAT	8 bits	Interrupt Status Register
0x04000202,3	IDE-RESET	16 bits	IDE Drive reset control
0x04000300	INTF-CNTL	8 bits	Interface Control Register
0x04000301	INTF-STAT	8 bits	Interface Status Register

INTF-CNTL Interface Control Register

The assignments of this register are as follows:

7	6–5	4	3	2	1	0
ROMO	INTL10	AHA\	DMA17U	NDRESET	ROM1	(not currently used)

ROM Enables the expansion card ROM BIOS and chooses its address

00 ROM disabled
01 ROM enabled at 0xc800
10 ROM enabled at 0xcc00
11 ROM enabled at 0xd800

INTL Determines the cards EISA interrupt level.

00 Interrupt 15
01 Interrupt 14
10 Interrupt 12
11 Interrupt 11

AHA\Indicates which type of interface DDA is currently emulating.

1 WD1003 task file register set
0 AHA-1542 register set

DMA17 Address bit 17 for DMA transfers.
NDRESET IDE Drive's RESET line

INTF-STAT—Interface Status Register

Assignments in this register are as follows:

Reg Addr	Indicates the last register read or written by the host see section 7.5 on page 16 for the exact mapping.
DF	Indicates the state of the AHA DF flag.
CDF	Indicates the state of the AHA CDF flag.
DHGT-PRES (two bits)	Indicates the type and presence of a daughtercard.

Reg Addr Indicates the last register read or written by the host see section 7.5 on page 16 for the exact mapping.
DF Indicates the state of the AHA DF flag.
CDF Indicates the state of the AHA CDF flag.
DHGT-PRES (two bits) Indicates the type and presence of a daughtercard.

Reg Addr Indicates the last register read or written by the host see section 7.5 on page 16 for the exact mapping.

DF Indicates the state of the AHA DF flag.

CDF Indicates the state of the AHA CDF flag.

DHGT-PRES (two bits) Indicates the type and presence of a daughtercard.

INT-STAT—Interrupt Status Register

7	6	5	4	3	2	1	0
TEOP_	DEOP_	DGHT_INT	H5INT	H4INT	H3INT	H2INT	H1INT

TEOP_ Indicates that the last burst terminated but was interrupted by a register cycle
DEOP_ Indicates that a task file PIO transfer has completed.
DGHT_INT Indicates that the daughter card needs servicing.
H5INT Indicates a drive interrupt from either master on drive connector 5.
H4INT Indicates a drive interrupt from either master on drive connector 4.
H3INT Indicates a drive interrupt from either master on drive connector 3.
H2INT Indicates a drive interrupt from either master on drive connector 2.
H1INT Indicates a drive interrupt from either master on drive connector 1.

IDE-RESET—IDE Reset Control Register

F	E	D	C	B	A	9	8	7	6
CLRINT1	SETMINT	NO_REG	CCDF	SDF	DRQA1	DRQA0	DRQen	RE2	RE1

CLRINT1	Clears the interrupt gate.
SETMINT	Sets the EISA interrupt described by INTL
CCDF	Clears the AHA CDF flag in the status register in AHA mode. Also used to clear the busy status in the status register in WD1003 mode. 0 Clear it 1 Don't clear it
SDF	Sets the AHA CDF flag in the status register in AHA mode. 0 Don't set it 1 Set it
NO_REG	Stops emulation register cycles from clearing the DMA counter (Commonly referred to as the NO_REG state.)
DRQA	Indicates which of the four supported lines is asserted 0 DRQ0 1 DRQ6 2 DRQ5 3 DRQ7
DRQen	Enables the deassertion of the DRQ lines
RE2, RE1, RE0	This gates (logical "1") or allows ("0") the DEOP signal to set busy.

EISA Side I/O Map

The EISA side I/O map consists of two regions which are implemented in two pieces of hardware:

The slot specific shared I/O registers in the BMIC at addresses 0xXC80—0xC9F.

The Adaptec AHA-154x register set which is emulated by dual porting the SRAM.

Interrupts

The interrupts in descending priority are:

BMIC,

Register Emulation,

IDE drives, and

System clock.

AHA Mode: AHA-1540 Register Set Emulation

The AHA-1540 register set consists of three eight bit registers. The base address of these registers can be set at addresses:

0x0334 which is the default

0x0330

0x0234

0x0230

0x0134

0x0130

The register set is:

Address	Read	Write
Base	Status Register	Control Register
Base + 1	Data In Register	Command Register
Base + 2	Interrupt Flags Register	Reserved

A write to any of the AHA1540 registers will generate an interrupt to processor **208**.

Status Register

7	6	5	4	3	2	1	0
STST	DIAGF	INIT	IDLE	CDF	DF	0	INVDCMD

STST	Self test in progress
DIAGF	Diagnostics failed
INIT	Mailbox initialization required
IDLE	Host adapter idle
CDF	Command/Data out port full - The host uses CDF to synchronize command and data transfers to the host adapter. An adapter command byte or an outbound parameter can be placed in the command/data out port when the port is empty, indicated by the CDF bit being zero. When a byte is placed in the command/data out port, the CDF bit is set to one and remains one until the host adapter has read the byte. When the CDF bit returns to zero, the next command or parameter byte can be placed in the port.

-continued

DF	Data in port full - The host uses DF to synchronize transfers of data from the host adapter to the host. When the DF bit is set to one, the host adapter has placed a byte in the data in port for the host to remove. When the host performs a read from the data in port address, the DF bit is cleared to zero and not set to one again until a new data byte has been placed in the data in port.
INVDCMD	Invalid host adapter command

IDLE Host adapter idle

CDF Command/Data out port full—The host uses CDF to synchronize command and data transfers to the host adapter. An adapter command byte or an outbound parameter can be placed in the command/data out port when the port is empty, indicated by the CDF bit being zero. When a byte is placed in the command/data out port, the CDF bit is set to one and remains one until the host adapter has read the byte. When the CDF bit returns to zero, the next command or parameter byte can be placed in the port.

DF Data in port full—The host uses DF to synchronize transfers of data from the host adapter to the host. When the DF bit is set to one, the host adapter has placed a byte in the data in port for the host to remove. When the host performs a read from the data in port address, the DF bit is cleared to zero and not set to one again until a new data byte has been placed in the data in port.

INVDCMD Invalid host adapter command

Control Register

Control Register

7	6	5	4	3	2	1	0
HRST	SRST	IRST	SCRST				

HRST	Hard reset
SRST	Soft reset
IRST	Interrupt reset
SCRST	SCSI bus reset

Interrupt Flags Register

7	6	5	4	3	2	1	0
Any	0	0	0	SCRD	HACC	MBOA	MBIF

Any	Any interrupt - a logical OR of the rest of the register
SCRD	SCSI reset detected
HACC	Host adapter command complete
MBOA	Mailbox out available
MBIF	mailbox in full

Software Interface

Following is additional information regarding the specific software implementation choices of the presently preferred embodiment. Voluminous source code for the actual specific implementation is included in application Ser. No. 07/809,452, filed Dec. 7, 1991, entitled "Disk Controller with Dynamic Sector Remapping" (DC-183), which has common ownership and common effective filing date with the present application, and which is hereby incorporated by reference.

Native Mode

There are three types of native commands: physical, logical and extended. Physical commands are commands that are addressed to physical drives. These bypass the logical drive engine inside DDA Logical commands are addressed to logical drives and therefore go through DDA's logical drive engine. It is intended that only diagnostics and specialized utilities use the physical command interface and that all standard drive I/O go through logical mode. Extended commands are the last type of native commands. These consist of operations that must be performed by the controller that are not directly associated with drive I/O.

Native mode uses the BMIC mailbox registers, the BMIC semaphore registers, and the BMIC doorbell registers. The BMIC mailbox registers occupy 16 contiguous bytes in the EISA slot-specific I/O address range. The host processor (386/486) may access these registers 8, 16, or 32 bits at a time. On startup, the host software must determine the slot number containing DDA in order to know the upper nibble of the I/O address.

The BMIC mailbox is partitioned into two areas: the inbound area occupying the first 12 bytes and the outbound area occupying the last 4 bytes. BMIC semaphore **0** is used to claim rights to modify the inbound area and BMIC semaphore **1** is used to claim rights to modify the outbound area. All native mode commands use the inbound area to submit requests to DDA. DDA only allocates BMIC semaphore **1** and only releases BMIC semaphore **0**. The host should only allocate BMIC semaphore **0** and only release BMIC semaphore **1**. DDA will hold BMIC semaphore **0** on logical commands only when it receives a request which collides with a request already in process. Host software is obligated to avoid this situation. DDA also holds BMIC semaphore **0** on physical and extended requests since they use the inbound area for both input and output.

The doorbell mechanism of the BMIC is used to distinguish between different types of native commands. To issue a native command, the host processor uses BMIC semaphore **0** to gain ownership of the BMIC mailbox area. A BMIC semaphore is defined to be unallocated if a value of **0** is read from the flag bit. Once the host processor gains ownership of the mailbox, it writes a data structure into the BMIC mailbox and write to the BMIC local doorbell. For logical commands, the host should write a value of 0x10 to the local doorbell register. For physical commands a value of 0x20 is used and for extended commands a value of 0x40 is used. When the host receives the completion interrupt, the EISA doorbell register will indicate which command completed using the values above.

Finally, there are actually three more types of native commands: the hard reset/configuration command (doorbell value 0x80), the soft reset command (doorbell value 0x80), and foreign commands (doorbell values 0x01, 0x02, and 0x04). The reset commands are processed in the native decoder and are given their own doorbell bits. The reason for this is that software that issues the DDA hard reset command (notably the EISA CMOS initialization program) has no looping capability so it cannot play by the semaphore rules. To issue a DDA reset command, the host processor writes the reset data into the inbound area, then writes a value of 0x80 (hard) or 0x80 (soft) to the local doorbell register.

BMIC interface

The doorbell definitions are:

NTV-FOREIGN1-DOORBELL	0x01
NTV FOREIGN2-DOORBELL	0x02
NTV-FOREIGN3-DOORBELL	0x04
NTV-SRESET-DOORBELL	0x08

-continued

NTV-LOGICAL-DOORBELL	0x10
NTV-PHYSICAL-DOORBELL	0x20
NTV-EXTENDED-DOORBELL	0x40
NTV-HRESET-DOORBELL	0x80

Physical Mode

The system of the presently preferred embodiment (the Dell Drive Array, or "DDA") presents to a host operating system disk drive abstractions called Composite Disk Drives. Composite Disk Drives support logical requests that range from "read" and "write" to "scatter read with holes". A Composite Disk Drive (CDD) is physically implemented using (besides the hardware and firmware in the controller) one or more physical disk drives. Thus with ten physical disk drives as illustrated in FIG. **4**, two or more CDD could be formed with controller **200** controlling each of the CDDs. As an abstraction, the CDD hides many aspects of the physical drives from the host operating system. In particular, the abstraction hides, and hence prevents access to, the actual physical disk drive hardware interface.

For setup, maintenance, and diagnostic purposes, there is a need to get closer to the physical disk drive interface than is allowed through the CDD abstraction. For example, when a physical disk drive (PDD) is "new" and not yet part of a CDD, a means is needed to test the PDD and to write configuration information onto the PDD. Even when a PDD is a part of a CDD, there is a need to test the PDD and perhaps write new configuration information onto the PDD. In addition to these straightforward needs, it turns out that there is a need to access the PDD interface in order to perform drive vendor specific functions. Since these functions are vendor specific and since vendors and these functions change over time, there is strong motivation to accommodate access to these functions without changing DDA firmware.

To meet these needs, DDA has a Physical Mode Programming Interface. This interface is not normally disclosed to DDA owners or users but is used by Dell's DDADIAG program and by the DDA specific portion of the EISA Configuration program.

Synchronization of Physical Mode (PM) Commands with Logical (CDD) Commands

Physical Mode commands may be issued by the host at any time, including periods where the host is also issuing logical CDD commands. PM commands must be able to run without disturbing (other than the obvious requirement to be running only one command per drive at a time) the operation of the CDD. When a PM command is received, PM checks to see if the physical drive specified is part of a CDD. If it is not, the command is run without regard to CDD interference. If the physical disk drive specified is part of a CDD, PM synchronizes the command with the CDD driver by submitting a PHYSICAL logical request to the CDD driver. When the PHYSICAL request reaches the head of the CDD request queue, the CDD driver "executes" it. Execution of the PHYSICAL command consists primarily and simply of calling the request's Return Function, which in this case, happens to be the core PM request driver. In other words, PM get's the CDD driver to run the PM command. Synchronization is obviously ensured.

In addition to the simple single command PM/CDD synchronization above, there is a multi-command synchronization mechanism that is part of and used with the primitive PM command set. When the host wants to run only PM commands on a disk drive for a period of greater than one command or wants to use the other primitive commands, the host will issue the BE_ GIN_PHYS_MODE_ONLY command. When the host is ready to allow CDD commands to resume, it issues the END_PHYS_MODE_ONLY command.

When PM receives the BEGIN_PHYS_MODE_ONLY command, as with other commands, it checks to see if the physical drive specified is part of a CDD. As with other commands, if the drive specified is not part of a CDD, the command is run directly. It the drive specified is part of a CDD, PM get's the CDD driver to run the command as a PHYSICAL command return function. BPMO increments a phys_mode_only counter associated with the physical drive. It also increments a phys_mode_only counter associated with the CDD. Synchronization is attained by having the CDD driver refuse to run any commands when its phys_mode_only counter is non-zero.

Since the CDD driver will refuse to run any commands when in phys_mode_only mode, a refinement needs to be mentioned here. When PM runs a command, it does so by handing it to the CDD driver only if 1) the specified drive is part of a CDD and 2) the associated CDD is not in phys_mode_only mode.

When PM receives the END_PHYS_MODE_ONLY command, it performs the same CDD checks as with other commands. In normal operation, the associated CDD, if any, will be in phys_mode_only mode and so the EPMO command will be run directly by PM. The EPMO command decrements the phys_mode_only counter associated with the specified physical drive and decrements the phys mode only counter associated with the associated CDD, if any. If the EPMO command causes a CDD's phys_mode_only counter to go to zero, the CDD is obviously no longer in phys_mode_only mode. At this point, the CDD driver may have logical CDD commands queued that it has refused to run. To ensure a timely restart of the CDD, PM must issue an innocuous command to the CDD driver but only after it is no longer in phys_mode_only mode. PM does this by issuing a PHYSICAL command to the CDD driver with an innocuous return function. The technical term for this action is "Thump", as in PM "thumps" the CDD driver.

Rudimentary PM Command Set

PM's rudimentary command set consists of a number of AT task file "like" commands that are still abstract like logical CDD commands in that they do not provide for direct access to the physical disk drive interface but are closer to the physical disk drive interface nevertheless. These commands are: READ, READ with no retries, READLONG, READLONG with no retries, IDENTIFY, READBUF, WRITE, WRITE with no retries, WRITELONG, WRITELONG with no retries, FORMAT, WRITEBUF, SETBUF, SEEK, RECAL, VERIFY, VERIFY with no retries, INIT, DIAG, READMULT, WRITEMULT, SETMULT, and RESET.

Primitive PM Command Set

PM's six primitive commands are provided through a rudimentary "EXTENDED" command, although that is an arbitrary implementation detail. The BPMO and EPMO primitive commands have already been discussed. The remaining four primitive commands provide the host almost direct contact with the physical disk drive hardware inter-

face. The ISSUE_CMD command writes host specified values to the physical disk drive's task file registers. The RETURN_STATUS command reads the physical disk drive's task file registers and returns the values to the host. The READ_DATA command reads the indicated amount of data from the disk drive's data register and returns the data to the host. The WRITE_DATA command writes the host provided data to the disk drive's data register.

With these four primitive commands, the host can perform almost all standard task file commands, and can perform any of the vendor unique commands that we are currently aware of. Standard commands that cannot be performed include READLONG and WRITELONG. In addition, access is not provided to the alternate status register, the drive address register or to the device control register and drive interrupts are not reflected to the host. These limitations could be overcome by adding primitive commands and should not be thought of as limiting the scope of this disclosure.

The PM commands are currently used to 1) enable spindle sync on the CONNER 3204F 200Meg drives and 2) to download firmware to the Maxtor LXT series drives.

Physical Commands

For physical commands, the inbound area is used for both input and output. This is done because physical commands provide more than 4 bytes of output information. Mailbox performance is not an issue for physical commands since command processing is relatively slow and synchronous inside the IDE driver anyway. The native command interface simply retains ownership of the inbound area for the duration of physical command processing and relinquishes ownership only after generating the host interrupt. Physical command input looks like:

Host Address	Mailbox	Parameter
?C90	10	command byte input
?C91	11	drive number
?C92	12	transfer count
?C93	13	unused
?C94–?C95	14–15	starting cylinder
?C96	16	starting head
?C97	17	starting sector
?C98–?C9B	18-1b	linear host address

Physical command output looks like:

Host Address	Mailbox	Parameter
?C90	10	error
?C91	11	drive number
?C92	12	transfer count
?C93	13	status
?C94–?C95	14–15	starting cylinder
?C96	16	starting head
?C97	17	starting sector
?C98–?C9B	18-1b	linear host address

Physical commands are basically the same as task file commands. The command bytes are the same as those of task file commands and the return fields reflect the register values of the physical drive on completion of the command.

Physical Command Bytes

Command	Value	Definition
AT-RECAL	0x10	recalibrate command
AT-READ	0x20	read sector
AT-READL	0x22	read sector long
AT-READNR	0x21	read sector no retries
AT-READLNR	0x23	read sector long no retries
AT-WRITE	0x30	write sector
AT-WRITEL	0x32	write sector long
AT-WRITENR	0x31	write sector no retries
AT-WRITELNR	0x33	write sector long no retries
AT-VERIFY	0x40	read verify
AT-VERIFYNR	0x41	read verify no retries
AT-FORMAT	0x50	format track
AT-SEEK	0x70	seek
AT-DIAG	0x90	perform diagnostics
AT-INIT	0x91	initialize drive parameters
AT-READBUF	0xE4	read sector buffer
AT-WRITEBUF	0xE8	write sector buffer
AT-IDENTIFY	0xEC	identify drive
AT-SETBUF	0xEF	set buffer mode
AT-READM	0xC4	read multiple
AT-WRITEM	0xC5	write multiple
AT-SETMULT	0xC6	set multiple code

Physical Status Bytes

Bit Name	Value	Definition
AT-STATUS-BUSY	0x80	drive is busy
AT-STATUS-DRDY	0x40	drive is ready
AT-STATUS-DWF	0x20	write fault
AT-STATUS-DSC	0x10	seek complete
AT-STATUS-DRQ	0x08	data request
AT-STATUS-CORR	0x04	correctable error
AT-STATUS-IDX	0x02	index pulse
AT-STATUS-ERR	0x01	uncorrectable error

Physical Error Bytes

Bit Name	Value	Definition
AT-ERROR-BBK	0x80	bad block found
AT-ERROR-UNC	0x40	uncorrectable error
AT-ERROR-IDNF	0x10	sector ID not found
AT-ERROR-TO	0x08	time out
AT-ERROR-ABRT	0x04	command aborted
AT-ERROR-TK0	0x02	track 0 not found
AT-ERROR-AMNF	0x01	address mark not found

Logical Commands

For logical commands, both the inbound and outbound areas are used. This allows DDA to complete one request while the host is submitting another. In addition, DDA has a simple mechanism. for allowing multiple logical commands to be processed, although logical commands must be submitted one at a time.

Logical command input looks like:

Host Address	Mailbox	Parameter
?C90	10	command byte input
?C91	11	drive number
?C92	12	transfer count
?C93	13	optional host request id

-continued

Host Address	Mailbox	Parameter
?C94–?C97	14–17	starting sector specification
?C98–?C9B	18-1b	host address specification

Logical command output looks like:

Host Address	Mailbox	Parameter
?C9C	1c	status
?C9D	1d	transfer count
?C9E	1e	host request id return
?C9F	1f	unused

Starting Sector Specification

Ordinarily, the starting sector is a zero-based absolute sector number. As an alternative, a BIOS INT-13 compatible input form is provided. In this form the 32-bit number is encoded as:

Host Address	Mailbox	Parameter
?C94–?C95	14–15	8086 register CX
?C96–?C97	16–17	8086 register DX

where host registers CX and DX are as they are received by BIOS INT-13. In order to use this alternate encoding, DDA must know the geometry used on the BIOS side through the native extended command NTV-SETPARM. This should be done at BIOS initialization for each logical disk. Bits **0–5** of the logical command byte are reserved for the actual logical command. Setting bit **6** of the command byte indicates to use the alternate starting sector specification.

Host Address Specification

Ordinarily, the host address is a 32-bit linear address. This is compatible with 32-bit software and 8086 real mode after segment:offset linearization. As an alternative, DDA will linearize the address for the host assuming the address is in real mode segrnent:offset form. In this form the 32-bit number is encoded as:

Host Address	Mailbox	Parameter
?C98–?C99	18–19	8086 register BX
?C9A–?C9B	1a–1b	8086 register ES

where host registers BX and ES are as they are received by BIOS INT-13. In order to use this alternate encoding, the host must be in real mode. Bits **0–5** of the logical command byte are reserved for the actual logical command. Setting bit 7 of the command byte indicates to use the alternate host address specification.

Logical Command Handles

For host software which would like to submit multiple logical commands to DDA, a unique ID must be chosen for each outstanding request. The DDA extended native command NTV-GETLCSIZE may be used to determine the ID range. The host software chooses the ID, or handle, and passes it in using the fourth byte of the inbound area. When DDA is finished with the request, it will return the handle in the third byte of the outbound area. The host software is responsible for choosing handles which are not currently in use. In the event that a duplicate handle is given to DDA, DDA holds the inbound request and the inbound semaphore until the previous request completes. This insures data integrity inside DDA at the cost of making the native interface semi-synchronous. For host software that chooses not to use the feature or cannot take advantage of it, simply ignore the handle field and process one request at a time. DDA reserves the right to process input requests out of sequence, though data integrity is always insured. In other words, DDA may reorder reads, but never writes unless it would be safe to do so.

Miscellaneous Logical Commands

Several logical commands don't actually perform drive I/O. These are NTV-READBUF, NTV-WRITEBUF, NTV-READCRC, NTV-WRITECRC, and NTV-IDENTIFY.

NTV-READBUF, NTV-WRITEBUF, NTV-READCRC, and NTV-WRITECRC are used to check DMA functionality. For these commands all arguments are ignored. DDA simply sets aside a 512 byte buffer to perform the diagnostic transfers. NTV-READCRC/NTV-WRITECRC are the same as NTV-READBUF/NTV-WRITEBUF, except the first 508 bytes are crc'ed and compared to the last 32 bits. This allows the DMA controller to be tested independently in each direction.

NTV-IDENTIFY ignores all but the drive field of the input arguments. It writes a single 512 byte buffer which contains the ntvIdentify structure associated with the composite drive.

Logical Command Byte

Command	Value	Definition
NTV-RECAL	0x00	recalibrate command
NTV-READ	0x01	read sector(s)
NTV-WRITE	0x02	write sector(s)
NTV-VERIFY	0x03	verify sector(s)
NTV-SEEK	0x04	seek
NTV-GUARDED	0x05	verify guard on sector(s)
NTV-READAHEAD	0x06	read sector(s) but no xfer
NTV-READBUF	0x07	native readbuffer diag
NTV-WRITEBUF	0x08	native writebuffer diag
NTV-WRITEVER	0x09	write and verify sector(s)
NTV-IDENTIFY	0x0A	logical unit info command
NTV-READCRC	0x0B	native readbuffer w/ crc
NTV-WRITECRC	0x0C	native writebuffer w/ crc
NTV-READLOG	0x0D	read first/next errlog entry

Logical Status Byte

Bit Name	Value	Definition
BADBLOCK	0x80	bad block found
UNCORECT	0x40	uncorrectable fault
WRITEFLT	0x20	write fault
IDNFOUND	0x10	sector id not found
CORRECT	0x08	correctable fault
ABORT	0x04	received abort from drive
TRACK0NF	0x02	track 0 not found
LTIMEOUT	0x01	logical drive timed out somehow
OK	0x00	no error

Logical Identify Command

The logical identify command returns a structure whose definition is:

```
typedef struct
{
    ulong totalSectors; /* total num of avail sectors */
    ushort    heads; /* logical heads */
    ushort    sectors; /* logical sectors per track */
    ushort    cylinders; /* logical cylinders available */
    ushort    pheads; /* physical heads */
    ushort    psectors; /* physical sectors per track */
    ushort    pcylinders; /* physical cylinders total */
    ushort    rcylinders; /* physical cylinders reserved */
    uchar maxTransfer; /* largest single xfer on lu */
    uchar multiple; /* largest single xfer on drive */
    ushort    dataBM; /* data drive bitmap */
    ushort    parityBM; /* parity drive bitmap */
    uchar configuredType; /* mirrored, guarded, etc */
    uchar type; /* mirrored, guarded, etc */
    uchar pupStatus; /* powerup status of the lun */
    uchar compSkew; /* num sectors volume */
    ulong patchAddr; /* abs sector of patch start */
    ulong errLogAddr; /* abs sector of errlog start */
    ulong errLogSectors; /* num sectors in errlog area */
    ulong numErrEvents; /* number error events logged */
    ulong numRemapped; /* number remapped stripes */
    ulong driveErrors[10]; /* number errors per drive */
    ulong firmwareRev; /* firmware revision number */
    ulong patchRev; /* patch revision number */
    uchar EmulMode; /* AHA or AT mode */
    uchar MaxReadAhead; /* number of sectors readahead */
    uchar PostWrites; /* post writes enabled? */
    uchar CacheEnabled; /* cache enabled? */
    ulong BmicBurstSize; /* largest DMA burst in bytes */
    uchar sourceRev[32]; /* rcs revision of source code */
} ntvIdentify;
```

Extended Commands

Like physical commands, extended commands use the inbound area for both input and output. This is done because some extended commands provide more than 4 bytes of output information. Extended commands are used to communicate configuration information between the host a DDA and perform no disk I/O. The native command interface simply retains ownership of the inbound area for the duration of extended command processing and relinquishes ownership only after generating the host interrupt.

Extended commands consist of both GET and SET instructions. GET instructions provide the host with information. SET instructions provide DDA with information. Each extended command must be individually described.

The following are the extended command enumerations:

Command	Value	Definition
NTV-DIAG	0x01	perform ctlr diagnostics
NTV-GETVERSION	0x02	get DDA version numbers
NTV-GETLCSIZE	0x03	get max logical cmd handle
NTV-GETPHYSCFG	0x04	get physical disk config
NTV-SETPARM	0x05	set parms of logical disk
NTV-GETNUMLU	0x06	get number of logical disks
NTV-GETLUCAP	0x07	get capacity of logical disk
NTV-SYNC	0x08	wait until DDA has finished
NTV-PUPSTAT	0x09	status of power up
NTV-DORESTORE	0x0a	begin restore process
NTV-PROGRESS	0x0b	give restore progress
NTV-DIAGREAD	0x0c	manf IDE port read
NTV-DIAGWRITE	0x0d	manf IDE port write
NTV-DATETIME	0x0e	set date and time
NTV-GETHWCFG	0x0f	get hardware configuration

The format of the mailboxes for each extended command is described below.

NTV-GETVERSION - Gets DDA firmware revision numbers

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90–?C91	DDA firmware major number
	?C92–?C93	DDA firmware minor number
	?C94–?C97	DDA firmware patch revision
	?C98–?C9B	DDA firmware part no.

NTV-GETLCSIZE - Gets number of logical command handles

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90	number of logical command handles
	?C91	max size of native S/G list
	?C92–?C93	suggested max command queue depth

NTV-GETPHYSCFG - Get physical drive configuration

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90–?C91	Good drive bitmap
	?C92–?C93	Bad drive bitmap
	?C94–?C95	Unreadable drive bitmap

NTV-GETNUMLU - Gets number of logical drives and the emulation mode

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90	number of logical drives
	?C91	0 = NONE 1 = AHA 2 = AT

NTV-GETLUCAP - Gets capacity of logical drive

Direction	Mailbox Address	Definition
input	?C90	command byte input
	?C91	logical disk number
output	?C90–?C93	total capacity
	?C94	logical geometry - heads
	?C95	logical geometry - sectors
	?C96–?C97	logical geometry - cylinders
	?C98–?C99	sectors per physical track
	?C9A	physical heads
	?C9B	current status

The status bytes returned are:

Status	Value	Definition
PUP-DEAD	0	controller died
PUP-OK	1	normal
PUP-NOTCONFIG	2	no configuration (virgin)
PUP-BADCONFIG	3	bad drive configuration
PUP-RECOVER	4	new drive - recovery possible
PUP-DF-CORR	5	drive failed - correctable
PUP-DF-UNCORR	6	drive failed - uncorrectable
PUP-NODRIVES	7	no drives attached
PUP-DRIVESADDED	8	more drives than expected
PUP-MAINTAIN	9	maintain mode

-continued

Status	Value	Definition
PUP-MANFMODE	10	manufacturing mode
PUP-NEW	11	new - needs remap generated

NTV-PUPSTAT - Controller status

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90	current overall status

The status bytes returned are the same as for NTV-GETLUCAP described above.

NTV-PROGRESS - Get restore progress

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90–?C93	total restored so far
	?C94–?C97	total to restore
	?C98–?C99	restorable composite drive bitmap

NTV-GETHWCFG - Get hardware configuration

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90	interrupt level (11, 12, 14, 15)
	?C91	DMA channel (0, 5, 6, 7)
	?C92	option ROM position (0–3)
	?C93	DDA major number (0–3)
	?C94–?C95	I/O Address
	?C96	0 = none; 1 = AHA; 2 = AT
	?C97	32-bit emulation control
	?C98	cache: 0 = off, 1 = on
	?C99	Write Strategy: 0 = sync, 1 = writethru, 2 = writeback

NTV-DIAG - Tells DDA to run internal diagnostics and return result

Direction	Mailbox Address	Definition
input	?C90	command byte input
output	?C90	completion status

NTV-SYNC—Tells DDA to finish all outstanding I/O

Direction	Mailbox Address	Definition
input	?C90	command byte input

NTV-SETPARM—set logical drive parameters.

Direction	Mailbox Address	Definition
input	?C90	command byte input
	?C91	drive
	?C92	heads
	?C93	sectors
	?C94–?C95	cylinders (unused)

NTV-DORESTORE—Tells DDA to begin the drive rebuild

Direction	Mailbox Address	Definition
input	?C90	command byte input
	?C91	0:background 1:foreground

NTV-DATETIME—Sets DDA clock to correct date and time.

Direction	Mailbox Address	Definition
input	?C90	command byte input
	?C94–?C97	days since January 1, 1980
	?C98–?C9B	seconds since 12:00 AM

Native SetParm Command

```
typedef struct
{
    uchar command;
    uchar drive;
    uchar heads;
    uchar spt;
    ushort cyls;
} ntvSetParm;
```

Supported SCSI Commands

Command	Value	Definition
SC-FORMAT	0 x 04	
SC-INQUIRY	0 x 12	
SC-MODE-SELECT	0 x 15	
SC-MODE-SENSE	0 x 1a	
SC-MEDIA-REM	0 x 1e	
SC-READ	0 x 08	Read block(s)
SC-READ-BUFFER	0 x 3c	
SC-READ-CAP	0 x 25	
SC-DEFECT-DATA	0 x 37	
SC-READ-EXT	0 x 28	
SC-REASSIGN-BLK	0 x 07	
SC-RELEASE	0 x 17	
SC-REQ-SENSE	0 x 03	
SC-RESERVE	0 x 16	
SC-REZERO-UNIT	0 x 01	
SC-SEEK	0 x 0b	
SC-SEEK-EXT	0 x 2b	
SC-SEND-DIAG	0 x 1d	
SC-START-STOP	0 x 1b	
SC-TEST-UNIT	0 x 00	
SC-VERIFY	0 x 2f	
SC-WRITE	0 x 0a	
SC-WRITE-BUFFER	0 x 3b	
SC-WRITE-EXT	0 x 2a	
SC-WRITE-VERIFY	0 x 2e	

Error Logging

DDA has the capability of logging errors to the disk. There is a native mode interface to retrieve error log information (see section above for details). The actual information logged may be varied. Alternatives include: For each drive error the following information is kept:

Time and date

Type of event

Drives reported status and error registers

Physical block and drive

Errors are logged for the following events:

Drive failures

Block remaps

Errors could potentially be logged for the following events:

Patches applied

Disk recover operations

Disk retries

Drive corrected errors

A further optional capability would allow the host can write to the error log.

DDA maintains separate logical and physical defect lists, as indicated by the firmware listed below.

Configuration Options

DDA can support any configuration of drives which can be cabled, providing the following rules are met:

An array logical drive can consist of at most five data drives.

A redundant logical drive can consist of at most four data drives and one parity drive.

A mirrored logical drive can consist of at most five data drives and five parity drives.

DDA supports many options, which are chosen using the EISA configuration utility described below.

The most important option is Adaptec 1540 emulation mode (AHA mode). However, emulation can be disabled, in which case only native mode is active.

Adaptec 1S40 Emulation Mode (mode AHA)

Options for this mode are:

I/O address: [330h,334h,230h,234h,130h,134h]

Interrupt Level: [11h,12h,14h,15h]

For a DDA configured as the primary (bootable) disk controller, the interrupt level should be set to 11h.

DMA channel: DMA request channel [0,5,6,7]

AutoRequest Sense: [YES,NO]

scsiID:

boardID:

specialOptions:

fwRevByte0:

fwRevByte1:

AHA mode can address up to 7 composite drives.

Native mode

Options for this mode are:

Interrupt Level: [11h,12h,14h,15h]

For a DDA configured as the primary (bootable) disk controller, the interrupt level should be set to 14h.

Native mode can address all composite drives.

An additional native mode implements the currently architected SCSI command to remap a block.

Option ROM address

The choices for this address are: C800h_CCOOh_ D800h_OFF. For DDA to be bootable, the Option ROM must be enabled.

Other options:

There are also several other options:

Posted-writes: Choice: [ON, OFF], Default: OFF.

When posted writes are enabled, DDA acknowledges a successful write after finishing the DMA transfer. After signaling completion, the actual write is immediately queued for writing. DDA cannot guarantee actual completion of a write or recover from a write error—thus introducing some risk when enabling this feature. Enabling posted writes can, however, significantly improve performance, especially for guarded composite drives.

Cache: Choice: [ON, OFF]; Default: ON.

DDA implements a cache in its sector buffer pool. For a variety of reasons, the DDA cache is not useful in some environments and may be disabled to improve performance. Performance is improved under many DOS-based benchmarks and therefore the default is ON.

DMA burst size: Choice: [256,1310721]; Default: 131070.

This option allows users to artificially limit the maximum EISA bus bandwidth that DDA will use. It is not clear that this option will be necessary; but for compatibility with unknown communications hardware, DDA's DMA capability may need to be artificially limited.

Maximum read-ahead size: Choice: [0,2551]; Default: 16.

This specifies the maximum number of sectors DDA will read beyond a given read request when possible. This option is only enabled when the cache is enabled.

Concurrent native mode requests: Choice: [16, 32, 64, 128]; Default: 16 w/emulation, 64 w/out emulation.

This specifies the total number concurrent requests that can be submitted to DDA through the native mode interface. For BIOS purposes this number may be small, but in the event native mode drivers are developed for a multi-tasking environment, this number should be set large. Each request consumes 48 bytes of DDA memory.

Further Modifications and Variations

The preferred embodiments may be modified in many ways while retaining one or more of the claimed innovative features, and the corresponding advantages. It will be recognized by those skilled in the art that the innovative concepts disclosed in the present application can be applied in a wide variety of contexts. Moreover, the preferred implementation can be modified in a tremendous variety of ways. Accordingly, it should be understood that the modifications and variations suggested below and above are merely illustrative. These examples may help to show some of the scope of the inventive concepts, but these examples do not nearly exhaust the full scope of variations in the disclosed novel concepts.

Of course, the specific components used are merely illustrative. For example, a different CPU chip, or multiple chips, or multiple CPUs can be used instead in alternative embodiments.

For another example, it is contemplated that in future product generations controller **100** would also provide a sorting of the queue of requests from the host for access to the array and if two or more requests involve close proximity sectors (which may imply close track locations on the disk drive platters), then controller **100** combines such requests into a single request for contiguous sector access but with data transfers only on the appropriate sectors. That is, this effectively caches the intervening sectors and immediately subsequent requests from the host frequently involve the data on these initially omitted sectors. Such combination of requests increases data transfer speed by reading/writing the data with limited read/write head seeks. Note that the operating system for the host may put pending requests into an elevator queue and automatically combine requests with close logical locations.

It is also contemplated that, in future drive array controllers, head scheduling algorithms can be added for further optimization of request queues.

It should also be noted that some, but not all, of the innovative teachings disclosed in the present application can be advantageously applied to more conventional disk controllers, or at least to controllers which do not control a composite disk drive. Others of the inventions may be applicable only to a redundant disk array, and not to other types of disk array or composite disk drive. Others of the inventions may be less advantageously applicable to a single disk array than to a composite disk drive, and/or may be most advantageously applicable only to a redundant disk array, but less advantageously applicable to other types of disk array or composite disk drive.

For example, the dual defect lists described above would merely reduce to a single defect list if adapted to a single-drive controller. Thus, the teachings disclosed above on this topic appear to be inapplicable to single-disk controllers.

For another example, the teachings disclosed above regarding dynamic read minimization is applicable to redundant disk arrays, but not to other types of array.

For another example, the teachings disclosed above regarding request fragmentation would be much less advantageous in an environment which did not use composite drives.

For another example, the teachings disclosed above regarding scatter/scatter transfer operations are perfectly applicable to single-disk controllers as well as to composite drives.

For another example, the teachings disclosed above regarding the physical mode of the controller would not be applicable to most single-disk applications, but could be advantageous (though less so than to disk arrays) in some single-disk configurations. For example, these teachings might be useful in a system in which the user data block startes at an offset address after a reserved system block, or in a system in which the disk is partitioned among multiple user data areas, or in a system where the disk controller is emulating a non-native disk interface, or for pass-through to a non-disk device (e.g. a tape drive).

For another example, the teachings disclosed above regarding the improved readahead strategy are perfectly well applicable to a single-disk controller.

For another example, the teachings disclosed above regarding periodic activation of physical drives provides an improved channel for error status to propagate upward, and may be useful in other multilevel disk interface architectures (e.g. if another layer were added to the interface). Alternatively, these teachings can even be generalized to apply to multilayer architectures for interfacing to non-disk physical devices of other types.

For another example, the teachings disclosed above regarding

For another example, the teachings disclosed above regarding

A further contemplated alternative, which was removed from the firmware of the presently preferred embodiment, is permuted addressing. In this alternative embodiment, the SRAM (and/or DRAM) is aliased into three different address spaces: Normal addressing—which works as expected; Permuted High addressing—which sends the low sixteen bits of a word to the address accessed, and the high 16 bits of the word to the address accessed with address bit 9 inverted; and Permuted Low addressing—which sends the high sixteen bits of a word to the address accessed, and the low 16 bits of the word to the address accessed with address bit 9 inverted, and address bits 2 and 3 incremented. That is, processor 208 can read 16 bits from each of two drives at once and get the result into one 32 bit register. This word consists however of data from two different sectors. Permuted addressing is provided so that in a single store the processor can write two halfwords into two different buffers which are separated by 512 bytes, or the length of a sector. There is no halfword swapper on the data bus, and so address bit 1 does not change on permuted accesses. To get this to work, one of the drives has to be read as a single device to get the track buffer pointers displaced in the two drives.

As will be recognized by those skilled in the art, the innovative concepts described in the present application can be modified and varied over a tremendous range of applications, and accordingly the scope of patented subject matter is not limited by any of the specific exemplary teachings given.

What is claimed is:

1. A method for defect tracking and avoidance in a computer system storing data on an array of disk drives comprising:

configuring the array into stripes, wherein each stripe includes a plurality of portions, wherein each portion of a stripe residing on a different disk drive of the array;

maintaining a defect list of logical addresses of stripes containing portions that are defective;

maintaining a defect list of physical addresses of defective portions; and

storing the defect list of physical addresses in a reserved area of the disk drive array whose location is accessible independent of any logical drive configuration.

2. The method of claim 1 further comprising:

searching the defect list of logical addresses in response to a request to access data stored in the array.

3. The method of claim 1 further comprising:

detecting an error on a portion of a first stripe, the portion of the first stripe on a first disk drive;

remapping the data stored on the first stripe to a second stripe in the array;

updating the defect list of logical addresses with the logical address of the first stripe;

updating the defect list of physical addresses with the physical address of that portion of the first stripe having the error.

4. The method of claim 3 further comprising:

reconstructing the data in that portion of the first stripe having the error from the data stored on other portions of the first stripe, wherein the remapping the data further includes writing the reconstructed data to a portion of the second stripe.

5. The method of claim 3 further comprising:

receiving a request to access the data stored on the first stripe;

searching the defect list of logical addresses;

accessing the data stored on the second stripe as a result of the searching the defect list of logical addresses.

6. The method of claim 5 further wherein the receiving the request to access data stored on the first stripe includes a request to access data stored on a third stripe, the method further comprising:

splitting the request as a result of the searching the defect list of logical addresses;

accessing the data stored on the third stripe.

7. The method of claim **3** further comprising:

replacing the first disk drive with a new first disk drive;

searching the defect list of physical addresses;

remapping the data stored on the second stripe, to a third stripe where a portion of the third stripe residing on the new first disk drive.

8. A computer system comprising:

an array of disk drives;

a controller connected to the array of disk drives;

a bus interface providing a communications link between the controller and a host computer, the host computer providing data via the bus interface to the controller for storage on the array of disk drives;

the controller including:

means for storing the data on the array in stripes, wherein each stripe includes a plurality of portions, wherein each portion of a stripe residing on a different disk drive of the array;

means for maintaining a defect list of logical addresses of stripes containing portions that are defective;

means for maintaining a defect list of physical addresses of defective portions: and

means for storing, the defect list of physical addresses in a reserved area of the array of disk drives whose location is accessible independent of any logical drive configuration.

9. The computer system of claim **8** wherein the controller is capable of receiving a request from the host computer to access data from the array, the controller further including:

means for searching the defect list of logical addresses in response to receiving the request.

10. The computer system of claim **8** wherein the controller further comprising:

means for detecting an error on a portion of a disk drive in the array.

11. The computer system of claim **10** further comprising:

means for remapping data stored on a first stripe to a second stripe in the array, the data being remapped in response to the controller detecting an error on a portion of the first stripe.

12. The computer system of claim **11** wherein the controller further comprising:

means for updating the defect list of logical addresses with the logical address of the first stripe in response to the data being remapped;

means for updating the defect list of physical addresses with the physical address of the portion of the first stripe having the error.

13. The computer system of claim **12** wherein:

the controller further including means for reconstructing the data in that portion of the first stripe having the error from the data stored on other portions of the first stripe;

the means for remapping the data further includes means for writing the reconstructed data to a portion of the second stripe.

14. The computer system of claim **13** wherein the controller further includes:

means for searching the defect list of logical addresses in response to a request from the host computer to access data from the first stripe;

means for accessing the data stored on the second stripe as a result of the search of the defect list of logical addresses.

15. The computer system of claim **14** wherein:

the request from the host computer includes request to access data from a third stripe;

the controller further includes means for splitting the request as a result of the search of the defect list of logical addresses.

16. The computer system of claim **8** wherein:

the means for maintaining the defect list of physical addresses includes means for storing the defect list of physical addresses on a first disk drive of the array.

17. The computer system of claim **16** wherein:

the controller further including means for maintaining a second defect list of physical addresses of defective portions, the means for maintaining the second defect list include means for storing the second defect list on a second disk drive of the array;

the second defect list being a defect list of physical addresses of defective portions of the second disk drive;

the defect list of physical addresses being a defect list of physical addresses of defective portions on the first disk drive.

18. The method of claim **1** wherein the logical defects list is maintained in a remap data structure.

19. The computer system of claim **8** wherein the logical defects list is maintained in a remap data structure.

* * * * *

US005968182A

United States Patent [19]

Chen et al.

[11] Patent Number: 5,968,182

[45] Date of Patent: Oct. 19, 1999

[54] **METHOD AND MEANS FOR UTILIZING DEVICE LONG BUSY RESPONSE FOR RESOLVING DETECTED ANOMALIES AT THE LOWEST LEVEL IN A HIERARCHICAL, DEMAND/RESPONSE STORAGE MANAGEMENT SUBSYSTEM**

[75] Inventors: **James C. Chen**, San Jose; **Julia Liu**, Sunnyvale; **Chan Y. Ng**, San Jose, all of Calif.; **William G. Sherman, II**, Tucson, Ariz.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[21] Appl. No.: **08/854,441**

[22] Filed: **May 12, 1997**

[51] **Int. Cl.**[6] **G06F 11/00**

[52] **U.S. Cl.** **714/5**; 714/42; 714/8; 714/48

[58] **Field of Search** 395/182.03, 182.04, 395/182.05, 182.06, 183.18, 185.01, 185.07, 846, 858, 185.1, 182.09, 183.01, 183.17; 371/37.7, 40.13, 40.15; 364/728.02

[56] **References Cited**

U.S. PATENT DOCUMENTS

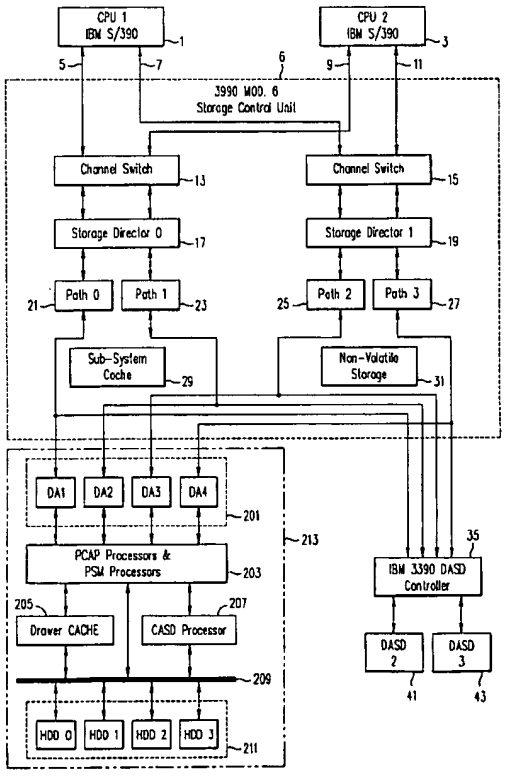
4,207,609	6/1980	Luiz et al.	395/858
4,761,785	8/1988	Clark et al.	371/51.1
5,191,584	3/1993	Anderson	371/51.1
5,210,860	5/1993	Pfeffer et al.	395/575
5,214,778	5/1993	Glider et al.	395/181
5,274,645	12/1993	Idleman et al.	395/182.04
5,274,794	12/1993	Ewing et al.	395/500
5,278,838	1/1994	Ng et al.	395/182.04
5,331,646	7/1994	Krueger et al.	371/40.1
5,367,669	11/1994	Holland et al.	395/575
5,504,859	4/1996	Gustafson et al.	395/182.09
5,550,543	8/1996	Chen et al.	341/94
5,617,425	4/1997	Anderson	371/10.2

Primary Examiner—Robert W. Beausoliel, Jr.
Assistant Examiner—Andy Nguyen
Attorney, Agent, or Firm—R. B. Brodie; E. E. Klein

[57] **ABSTRACT**

A method and means within a hierarchical, demand/response DASD subsystem of the passive fault management type in which, upon the occurrence of fault, error, or erasure, a long device busy signal of finite duration is provided to a host CPU. Any DASD storage device subject to the anomaly is isolated from any host inquiry during this interval. These measures permit retry or other recovery procedures to be implemented transparent to the host and the executing application. This avoids premature declarations of faults, errors, or erasures and consequent host application aborts and other catastrophic measures. If the detected anomaly is not resolved within the allotted time, then other data recovery procedures can be invoked including device reset, the status reported to the host, and the next request processed.

6 Claims, 4 Drawing Sheets



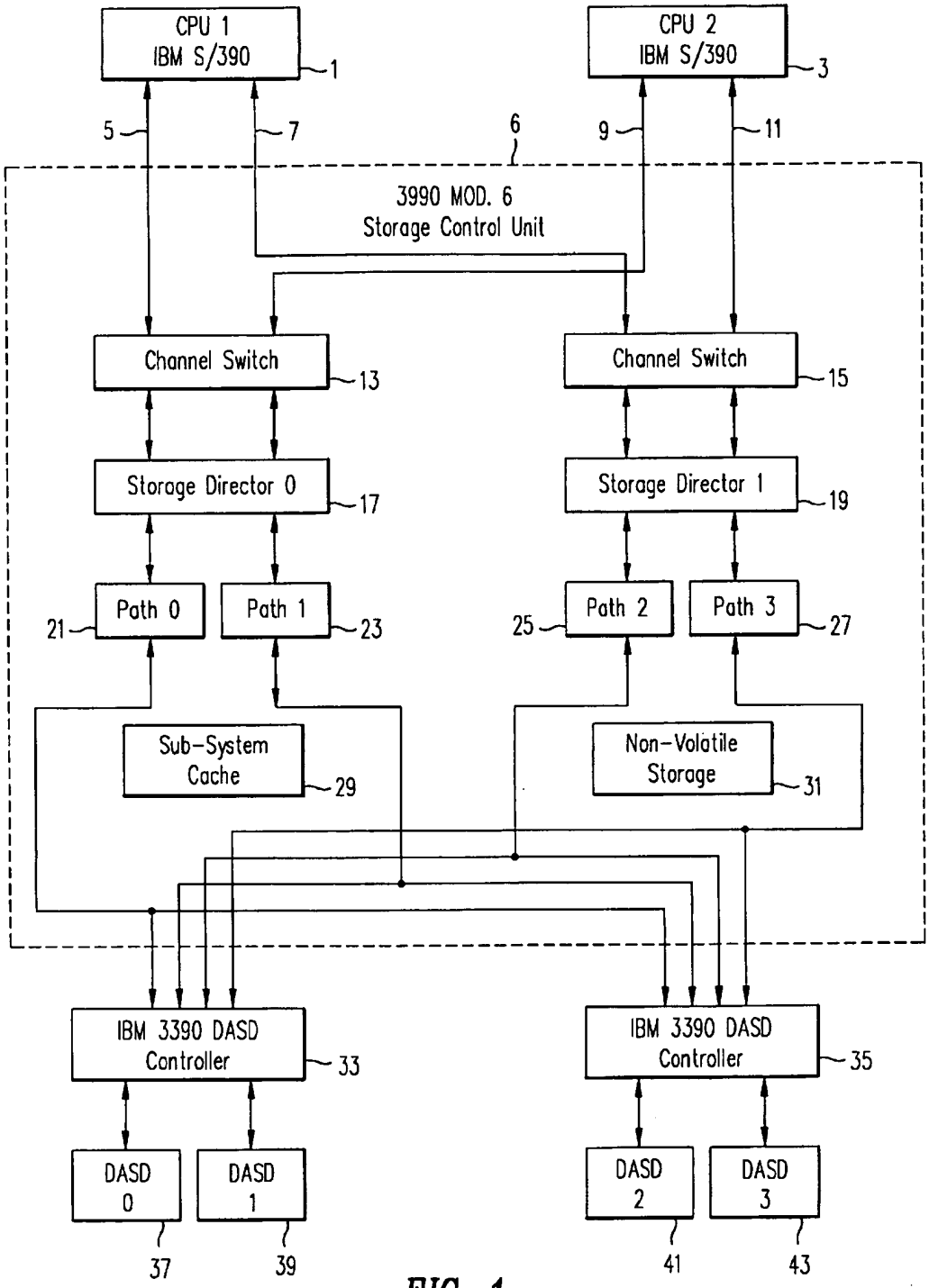

CPU 1
IBM S/390
1
CPU 2
IBM S/390
3
5
7
6
9
11
3990 MOD. 6
Storage Control Unit
Channel Switch
13
Channel Switch
15
Storage Director 0
17
Storage Director 1
19
Path 0
21
Path 1
23
Path 2
25
Path 3
27
Sub-System
Cache
29
Non-Volatile
Storage
31
IBM 3390 DASD
Controller
33
IBM 3390 DASD
Controller
35
DASD
0
DASD
1
37
39
DASD
2
DASD
3
41
43

FIG. 1

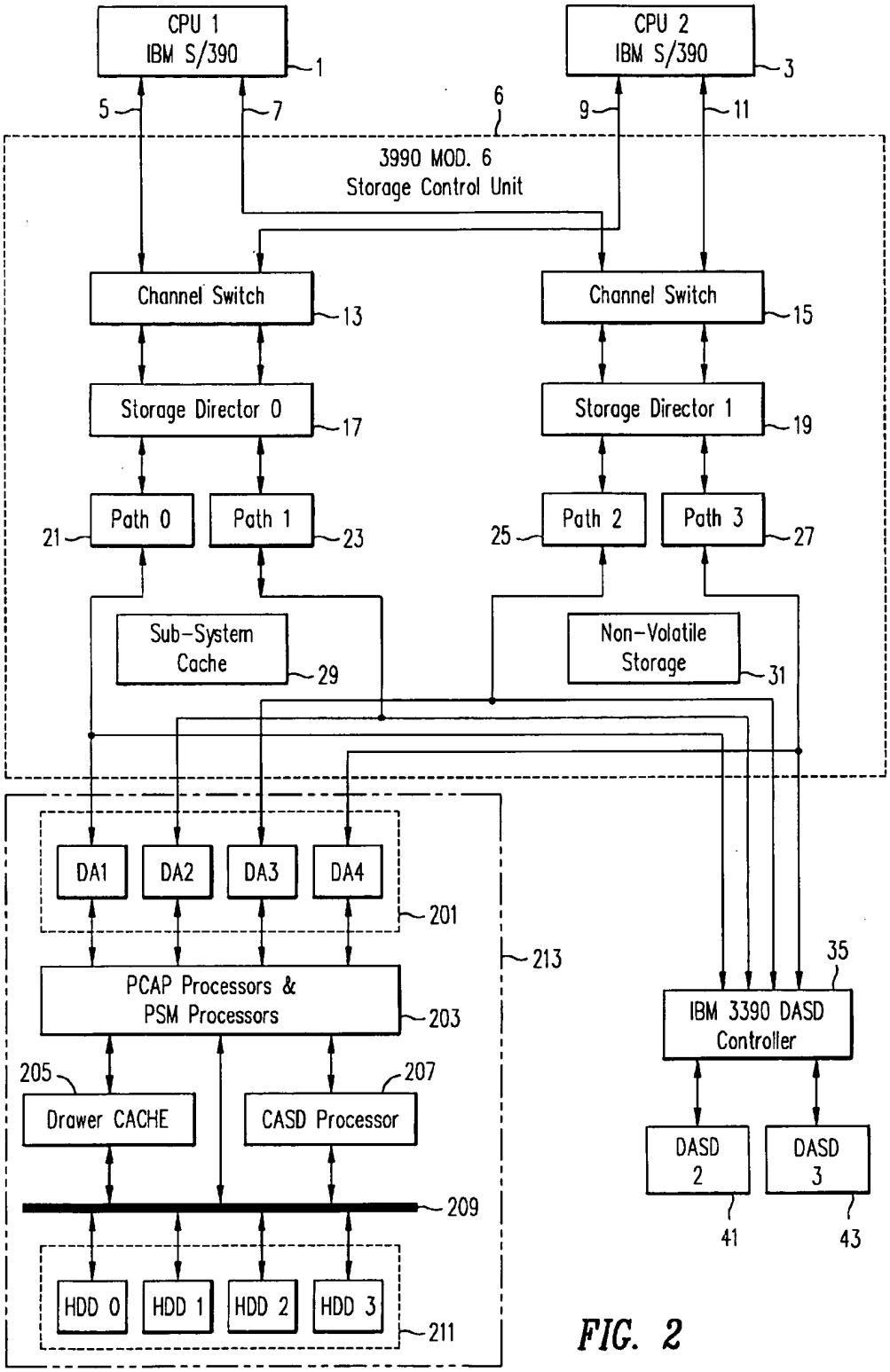

CPU 1
IBM S/390
1
CPU 2
IBM S/390
3
5
7
6
9
11
3990 MOD. 6
Storage Control Unit
Channel Switch
13
Channel Switch
15
Storage Director 0
17
Storage Director 1
19
21
Path 0
Path 1
23
25
Path 2
Path 3
27
Sub-System
Cache
29
Non-Volatile
Storage
31
DA1
DA2
DA3
DA4
201
213
PCAP Processors &
PSM Processors
203
35
IBM 3390 DASD
Controller
205
Drawer CACHE
207
CASD Processor
DASD
2
DASD
3
41
43
209
HDD 0
HDD 1
HDD 2
HDD 3
211

FIG. 2

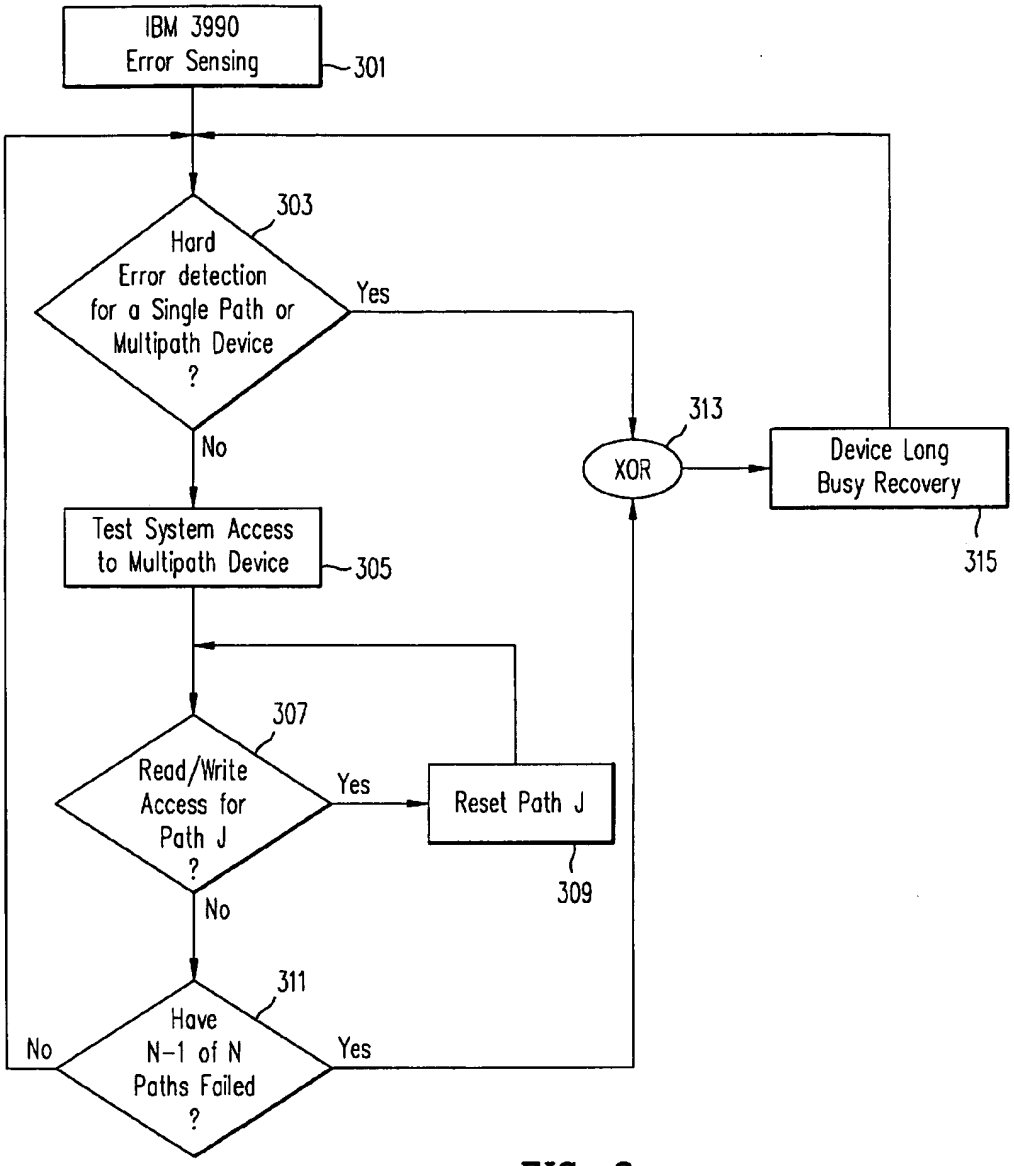

IBM 3990
Error Sensing
301
303
Hard
Error detection
for a Single Path or
Multipath Device
?
Yes
No
313
XOR
Device Long
Busy Recovery
315
Test System Access
to Multipath Device
305
307
Read/Write
Access for
Path J
?
Yes
Reset Path J
309
No
311
Have
N-1 of N
Paths Failed
?
No
Yes

FIG. 3

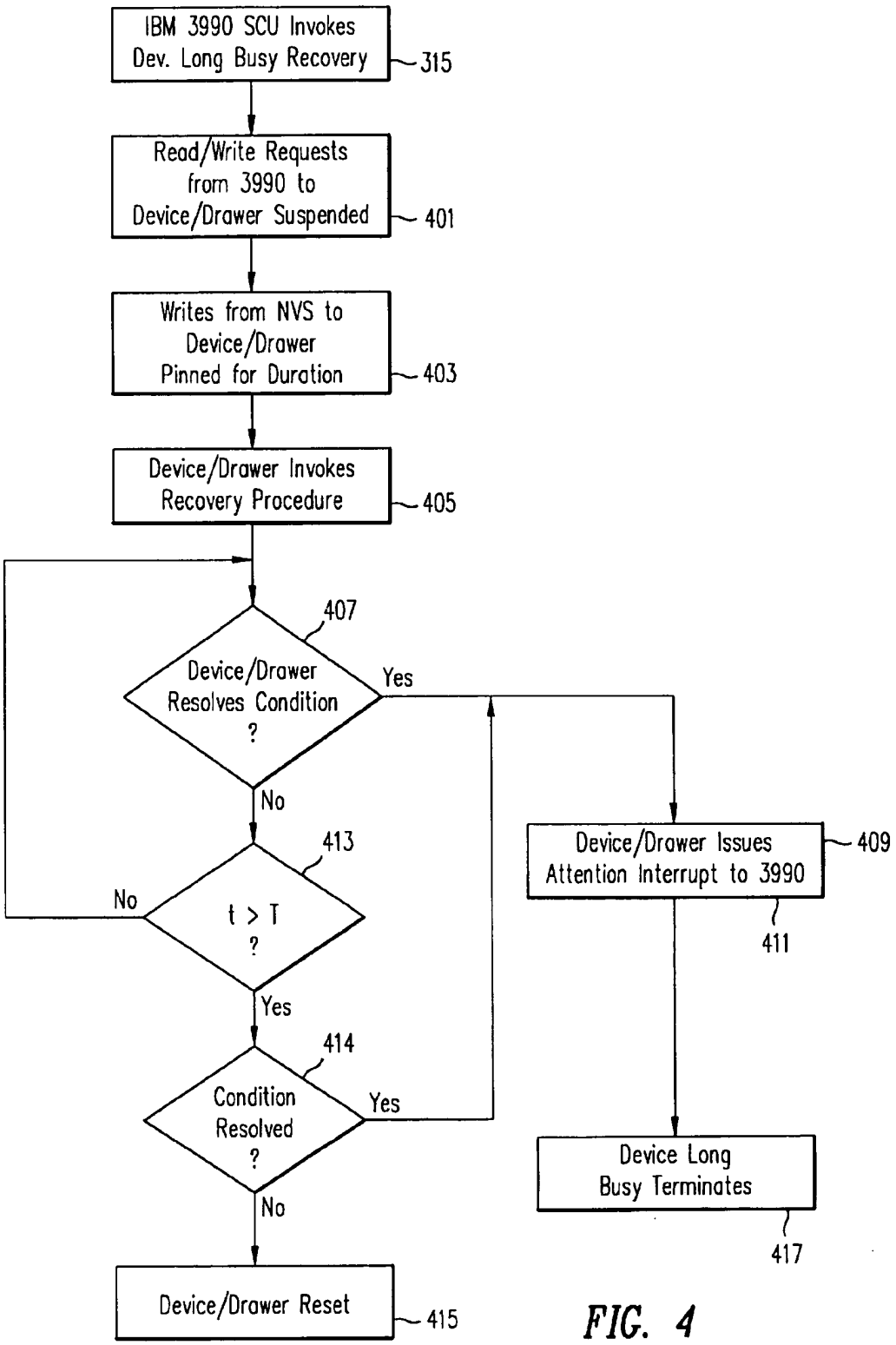

IBM 3990 SCU Invokes Dev. Long Busy Recovery
315
Read/Write Requests from 3990 to Device/Drawer Suspended
401
Writes from NVS to Device/Drawer Pinned for Duration
403
Device/Drawer Invokes Recovery Procedure
405
407
Device/Drawer Resolves Condition ?
Yes
No
413
t > T ?
No
Yes
414
Condition Resolved ?
Yes
No
Device/Drawer Reset
415
Device/Drawer Issues Attention Interrupt to 3990
409
411
Device Long Busy Terminates
417

FIG. 4

METHOD AND MEANS FOR UTILIZING DEVICE LONG BUSY RESPONSE FOR RESOLVING DETECTED ANOMALIES AT THE LOWEST LEVEL IN A HIERARCHICAL, DEMAND/RESPONSE STORAGE MANAGEMENT SUBSYSTEM

FIELD OF THE INVENTION

This invention relates to data recovery procedures in hierarchical, demand/response direct access storage device (DASD) subsystems and, more particularly, for managing status reporting to a host operating system as the attached DASD subsystem resolves detected anomalies.

DESCRIPTION OF RELATED ART

The disclosure is initiated with definitions relating to fault and failure, and continues with a brief discussion of hierarchical storage management systems. The section ends with discussions of the prior art modes of managing faults and failures in storage subsystems actively and passively where nesting failures upward can prematurely cause applications at hosts to abort.

Nomenclature of Faults, Failures, and Fault Tolerance

It is to be appreciated that in an information handling system, a "fault" is defined as a malfunction due to any one of several causes. In this regard, a fault may occur on either a transient, intermittent, or permanent basis. Also, a fault may be classified as either a "fail-silent fault" or a "Byzantine fault". Relatedly, a fail-silent fault is that type of malfunction wherein a component just terminates its performance, while a Byzantine fault is an undetected fault condition caused by hardware, software, or both. Technically, a system is said to "fail" when its behavior or activities do not conform to a specification. The same may be said for a "subsystem failure" or a "device failure".

"Fault tolerance" is the degree to which an information handling system, subsystem, device, or component can continue to operate notwithstanding the occurrence of one or more faults or failures. Fault tolerance is attained through the use of information, time, and physical redundancy.

Information redundancy uses additional information to detect, correct, or derive a bounded maximum of information in error, erasure, or unavailability. Time redundancy involves repeating actions otherwise incomplete without altering the system state. One example of time redundancy is "atomic transactions". An atomic transaction comprises a series of steps invoked by a process such that any interruption or failure to complete the series causes the system to return to its prior information state. Lastly, physical redundancy involves replacement of one portion of a physical computing, storage, or control layer with its performance clone.

Parenthetically, in this specification, the term "synchronous system" will be taken to mean a system having the property of always responding to a message within a known finite bound (T seconds). This includes time to process n repeat requests. Also, the term's "disk storage device", "direct access storage device", and the acronym DASD are used synonymously.

Aspects of Hierarchical Demand/Response Storage Subsystems and RAID 5 Arrays

In the period spanning 1970 through 1985, IBM developed large-scale multiprogramming, multitasking computers, S/360 and S/370 running under an MVS operating system. A description of the architecture and that of the attached storage subsystem may be found in Luiz et al., U.S. Pat. No. 4,207,609, "Method and Means for Path Independent Device Reservation and Reconnection in a Multi-CPU and Shared Device Access System", issued Jun. 10, 1980. Such systems were of the hierarchical and demand/responsive type. That is, an application running on the CPU would initiate read and write calls to the operating system. These calls were, in turn, passed to an input/output processor or its virtual equivalent (called a channel) within the CPU. The read or write requests and related accessing information would be passed to an external storage subsystem. The subsystem would responsively give only status (availability, completion, and fault) and pass the requested data to or from the CPU.

The architecture of hierarchical demand/response storage subsystems such as the IBM 3990/3390 Model 6 and the EMC Symmetrix 5500 is organized around a large cache with a DASD-based backing store. This means that read requests are satisfied from the cache. Otherwise, the data satisfying those requests are staged up from the DASDs to the cache. Write updates result in data being sent from the CPU to the cache or to a separate nonvolatile store (NVS), or both. This is the case with the IBM 3990 Model 6. The NVS stored data is then destaged or written out to the DASDs on a batched basis asynchronous to processing the write requests. The term "demand/response" connotes that a new request will not be accepted from a higher echelon until the last request is satisfied by a lower echelon, and a positive indication is made by the lower to the higher echelon.

In order to minimize reprogramming costs, applications executing on a CPU (S/390) and the attendant operating system (MVS) should communicate with invariant external storage architecture even though some components may change. Relatedly, the view of storage associated with an MVS operating system requires that data be variable length formatted (CKD) and stored on an external subsystem of attached disk drives (IBM 3390) at addresses identified by their disk drive cylinder, head, and sector location (CCHHSS). Requested variable length formatted data is staged and destaged between the CPU and the storage subsystem as so many IBM 3390 disk drive tracks worth of information.

It well appreciated that an improved disk storage facility can be attached to a subsystem if the new facility is emulation compatible with the unit it has replaced. Thus, a RAID 5 storage array of small disk drives can be substituted for a large disk drive provided there is electrical and logical interface compatibility. Illustratively, the IBM 3990 Model 6 storage control unit can attach an IBM 9394 RAID 5 array DASD and interact with it as if it were several IBM 3390 large disk drives. Data is staged and destaged to and from the RAID 5 array formatted as CKD formatted 3390 disk drive tracks. The RAID 5 array in turn will reformat the tracks as one or more fixed-block formatted strings and write them out to disk.

Active Fault Management

An active strategy in anticipation of fault, failure, and error would be to continuously monitor all data handling and component performance. Indeed, such systems are described in Glider et al., U.S. Pat. No. 5,214,778, "Resource Management in a Multiple Resource System", issued May 25, 1993, and in Idleman et. al., U.S. Pat. No. 5,274,645, "Disk Array System", issued Dec. 28, 1993.

Glider discloses a method and means for managing both subsystem control code (CC) and an active fault management system (FMS) competing for a disk-based storage subsystem resource access. In Glider, a subsystem uses resource availability states as semaphores to ensure serial-

ization where the FMS and the CC compete for access to the same resource. Significantly, the subsystem requires continuous availability monitoring for all resources for any changes.

Idleman describes a storage subsystem having a pair of two-level, cross-connected controllers providing a first and second failure-independent path to each DASD in a plurality of RAID 3 arrays of DASDs. Data is "striped" to support a parallel read or a parallel write across N+P+Q DASDs, where P and Q are redundancy bytes calculated over the N data bytes. That is, data is moved (parallel read or write) between controllers and a RAID 3 array using on-the-fly transverse redundancy error detection/correction and any-to-any switching between N+P+Q sources and sinks. As with prior art RAID 3, the redundancy blocks are bound to designated redundancy DASDs.

Passive Fault Management

In contrast to the Glider and Idleman references, a passive strategy can be used in a hierarchical, demand-responsive DASD storage subsystem exemplified by the IBM 3990/3390 Model 6. In a word, rather than hunt for fault or failure, the fault management is reactive. That is, the storage subsystem system relies on the presence of at least two failure-independent paths to a data storage location and the invocation of data recovery procedures (DRPs). The DRPs are invoked only upon the detection of error, fault, or failure in signals and data as they are read back or staged from a storage location.

Illustratively, a distorted modulated signal readback from a DASD track over a multibyte extent might cause a pointer to be generated at a signal processing level. It might also appear as a nonzero syndrome set at the ECC digital string reading level. At this point, an FMS would invoke DRPs to resolve the situation. Recovery actions might assume any one of a set of nested causes. The recovery actions themselves might range from a repetition of the read operation with or without a performance adjustment. For example, if the track/head misregistration was an assumed cause, then adjusting the head position relative to the track might be required. On the other hand, if thermal asperities were the assumed burst error cause, then ECC recovery from the syndrome set and the generated pointer might be the DRP of choice, etc.

Subsystem Complexity and Premature Termination of Recovery Actions

Where a hierarchical demand/response storage system attaches one or more RAID 5 DASD arrays in addition or instead of conventional DASDs, the likelihood of a RAID 5 array becoming incapacitated by a single storage element (HDD) failure resulting in system failure is remote. This derives from the fact that RAID 5 arrays have sufficient information and physical redundancy to fault tolerate at least one failure. This is also the case for even RAID 1 (mirrored pairs of IBM 3390 DASDs) and RAID 3 or RAID 4 array configurations.

In RAID 5 as described in Clark et al., U.S. Pat. No. 4,761,785, "Parity Spreading to Enhance Storage Access", issued Aug. 2, 1988, a parity group of n-1 fixed-size data blocks plus a parity block are written over n DASDs. The blocks of the parity groups are spread such that no single DASD has two blocks from the same parity group and no DASD has all of the parity blocks written thereon. In the event that a single DASD should fail, then the RAID 5 array can laboriously recover data from a referenced parity group by logically combining n-1 blocks from the remaining DASDs. Any additional DASD failure would result in a permanent failure for the array. Thus, restoration of both fault tolerance and reasonable response time requires rebuilding the data stored on the failed DASD and writing out to a spare DASD within the array. But the time required for rebuilding data on a spare varies under conditions of load on the remaining DASDs. This fact is well articulated in Ng et. al., U.S. Pat. No. 5,278,838, "Recovery from Errors in a Redundant Array of Disk Drives", issued Jan. 11, 1994.

Passive fault management has heretofore been designed to resolve well-defined faults or errors within relatively narrow bounds. For instance, if j repeated read accesses of a given DASD yields j repeated ECC errors over a variety of DRPs, then the DASD may be declared dead, i.e., treated as a failure. However, complex devices such as a RAID 5 array of small DASDs substituting for a single large DASD or admixed with them is unlikely to appear to the host or 3990 SCU as a hard disk failure. This means that the inflexible mode of status reporting and handling is more likely to result in frequent and premature termination of host-level applications. These are a subsystem reporting a device as having failed when it in fact did not, or correlatively reporting a device or operation as being successful when in fact it had either failed or was aborted.

SUMMARY OF THE INVENTION

It is accordingly an object of this invention to devise a method and means for flexibly scheduling the report to a host CPU of fault and error conditions detected in an attached hierarchical demand/responsive storage subsystem in order to minimize premature terminations of applications and other host-based catastrophic actions.

It is a related object that said method and means facilitate such flexible scheduling in a storage subsystem having a diverse attachment of storage elements or devices such as RAID arrays and stand-alone DASDs.

It is yet another object that said method and means operably perform even where the subsystem has significantly different ways of responding to error, fault, and failure conditions.

It was unexpectedly observed that if the subsystem, upon the occurrence of fault or failure, provided a long device busy signal to the host for up to a finite maximum duration and isolated the storage device from any host inquiry, then the variable duration data recovery procedures executed at the subsystem and device levels, especially those involving RAID 5 rebuild, could be executed. This would avoid premature declarations of hard faults, failures, and errors.

Restated, the foregoing objects are believed satisfied by a method and means for detecting and correcting a defective operating state or condition of a hierarchical demand/responsive storage subsystem attaching a host CPU. The subsystem includes a plurality of cyclic, tracked storage devices, an interrupt-driven, task-switched control logic, and circuits responsive to the control logic for forming at least one path of a set of paths coupling the host to at least one device. The host enqueues one or more read and write requests against the subsystem. Responsively, the subsystem control logic interprets each request and establishes a path to an addressed storage device.

The method steps of the invention include detecting any anomaly in the read back or staging of data from the device and executing a retry of the counterpart request by active or passive querying of said addressed device. In the event that the detected anomaly persists, a long busy status signal is presented to the host CPU by the control logic. In this regard, the long busy signal is an indication that the counterpart request has yet to be completed by the subsystem.

Next, access to the device is inhibited by the control logic for no more than a predetermined time interval. The method then ascertains whether the inhibited device has returned to an operational state. In the event the detected anomaly is resolved, an attention interrupt is set in the control logic by the device and the device long busy signal is terminated in the host CPU by the control logic. In the event that the time interval has been exceeded and the anomaly is not resolved, one or more data recovery procedures are invoked, including resetting the device by the control logic. Since the device has been driven into a final state, its status is then reported to the host CPU and the next request processed.

BRIEF DESCRIPTION OF THE DRAWING

FIG. **1** shows a logical block diagram of an IBM 3990/3390 illustrative of a hierarchical, demand/responsive storage subsystem.

FIG. **2** depicts the subsystem of FIG. **1** but is modified to set out the attachment of a RAID 5 DASD array as a logical 3390 DASD in addition to the attachment of real 3390 DASDs.

FIG. **3** illustrates the method of the invention as it initially responds to the detection of any anomaly in the read back or staging of data from a storage device whether a single large device or as the logical equivalent formed from an array of small devices.

FIG. **4** sets forth the method of the invention after a determination that the error is not one correctable after retry and only one path to the device path is available.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring now to FIG. **1**, there is shown a functional block diagram depiction of the IBM 3990/3390 Disk Storage Subsystem exemplifying a host-attached, hierarchical, demand/response storage subsystem. This subsystem is shown driven from first and second multiprogramming, multitasking hosts CPU **1** and **3**, such as an IBM System/390 running under the IBM MVS operating system. The subsystem is designed such that data stored on any of the DASD storage devices **37, 39, 41,** and **43** can be accessed over any one of at least two failure-independent paths from either one of the CPU's **1** or **3**. The system as shown provides four failure-independent paths. Illustratively, data on devices **37** or **39** can be reached via 3390 controller **33** over any one of paths **21, 23, 25,** or **27**. The same holds for data stored on devices **41** or **43** via controller **35**. A full description of this principle is to be found in the aforementioned U.S. Pat. No. 4,207,609, herein incorporated by reference.

The 3990 storage control unit consists of at least two storage directors **17** and **19**. These are microprocessors and attendant local memory and related circuitry (not shown) for interpreting control information and data from the CPUs, establishing logical and physical paths to the storage devices, and managing fault and data recovery at the subsystem level. The read and write transfer directions are separately tuned. That is, read referencing is first made to cache **29**, and read misses causes data tracks to be staged from the devices as backing stores. Write referencing either as a format write or an update write is made in the form of track transfers from the host to a nonvolatile store **31**. From NVS **31**, it is destaged to the devices through their sundry controllers.

Typically, an application executing on a host **1** or **3** requests to read a file, write a file, or update a file. These files are ordinarily stored on a large bulk 3990/3390 DASD storage subsystem **6**. The MVS host (S/390) is responsive to any read or write call from the application by invoking an access method. An access method, such as VSAM, is a portion of the OS for forming an encapsulated message containing any requested action. This message is sent to an input/output (I/O) portion of the host, and ultimately the storage subsystem. Typically, the message includes the storage action desired, the storage location, and the data object and descriptor, if any. This "message" is turned over to a virtual processor (denominated a logical channel). The function of the logical channel is to send the message to the storage subsystem over a physical path connection (channels **5, 7, 9, 11**). The storage subsystem control logic (director **17** or **19**) then interprets the commands. First, a path to the designated storage device is established and passes the interpreted/accessing commands and data object to the storage device location on a real time or deferred basis. The sequence of commands is denominated "channel command words" (CCWs). It should be appreciated that the storage device may be either "logical" or "real". If the device is "logical", then device logic at the interface will map the access commands and the data object into a form consistent with the arrangement of real devices. Thus, a RAID 5 array of small DASDs substitutes for one or more IBM 3390 large DASDs.

The "access method" portion of the MVS operating system, when processing data objects in the form of variable length ECKD records, also will ascertain either a "new address" or an old (update in place) address. The access method assumes that external storage includes actual physical DASDs, etc. devices. It generates addresses on a DASD device, cylinder, head, and record (CCHHRR) basis. Significantly, the data objects are ordinarily aggregated on a 3380/3390 DASD track basis. That is, when an application requests one or more records, the access method determines what would be an efficient unit of staging, i.e., record staging or track staging between the S/390 and the 3990 SCU. Accordingly, the access method modifies the CCW chain and address extent occasionally from a track to a record. In turn, the logical channel will cause a string of CCWs, together with "track-formatted" data, to be destaged to a 3990 storage control unit (SCU). An IBM 3990 storage control unit (SCU) "interprets" the CCWs and batches the writes in the nonvolatile store **31** (NV write buffer) for later destaging to one or more 3390 logical or physical DASDs **37, 39, 41, 43**. If a track is written out to a real 3390 DASD, then it will perform ECC processing as discussed subsequently. Originally, an access method comprised a set of protocols for moving data between a host main memory and physical input/output devices. However, today it is merely a mapping to a logical view of storage, some of which may be physical storage.

Referring now to FIG. **2**, there is depicted the subsystem of FIG. **1** but modified to set out the attachment of a RAID 5 DASD array **213** as a logical 3390 DASD, in addition to the attachment of real 3390 DASDs. In this regard, the IBM 3990 SCU Model 6 (FIG. **2/6**) utilizes a large cache (up to 2 gigabytes) (FIG. **2/29**). The data is always staged and destaged in the form of 3380/3390 tracks. This occurs when staging data between a plurality of logical (FIG. **2/213**) or real 3390 DASDs (FIG. **2/35, 41, 43**) and the 3990 cache (FIG. **2/29**) and destaging data between an NV write buffer (FIG. **2/31**) and the logical or real 3390 DASDs.

When track-formatted data is written out to the DASDs at the physical device, an ECC check byte is calculated over any destaged tracks and stored with the track. Upon any

subsequent read access, an ECC calculation over the staged tracks is again made and a comparison match between the stored values and the calculated values. Any mismatch is indicative of error. Restated, upon read back or staging of the data from a DASD, detection of any nonzero syndrome is an indication of random or burst error in the data.

Referring again to FIG. **2**, there is depicted a RAID 5 array **213** of small DASDs **211** attached to the control logic **17, 19** of the IBM 3990 storage control unit **6** over the plurality of paths **21, 23, 25**, and **27** via device adapters (DAs) **201**. One implementation of RAID 5 arrays is to be found in the IBM RAMAC Array DASD attaching one or more Enterprise System (S/390) ECKD channels through an IBM 3990 Model 3 or 6 storage control unit. The RAMAC Array DASD comprises a rack with a capacity between 2 to 16 drawers. Each drawer **213** includes four disk drives HDD0–HDD3, cooling fans, control processor **207**, ancillary processors **203**, and a nonvolatile drawer cache **205**. It is configured as a track staging/destaging to three DASDs' worth of data space and one DASD's worth of parity in a RAID 5 DASD array. Each drawer emulates between two to eight IBM 3390 Model 3 volumes.

Functionally, the DAs **201** provide electrical and signal coupling between the control logic **17** and **19** and one or more RAID 5 drawers. As tracks are staged and destaged through this interface, they are converted from variable length CKD format to fixed-block length FBA format by the ancillary processors **203**. In this regard, drawer cache **205** is the primary assembly and disassembly point for the blocking and reblocking of data, the computation of a parity block, and the reconstruction of blocks from an unavailable array of DASDs. In this embodiment, three DASDs are used for storing parity groups, and the fourth DASD operates as a hot spare. If a dynamic (hot) sparing feature is used, then the spare must be defined or configured a' priori. Space among the three operational arrays is distributed such that there exists two DASDs' worth of data space and one DASD's worth of parity space. It should be pointed out that the HDDs **211**, the cache **205**, and the processors **203** and **207** communicate over an SCSI-managed bus **209**. Thus, the accessing and movement of data across the bus between the HDDs **211** and the cache **205** is closer to an asynchronous message-type interface.

Since passive fault management is used, it should be pointed out that ECC correction is applied only to data as a serial stream read or staged from a given array storage device. The parity block is used only in recovery mode to reconstruct data from an unavailable or failed one of the array DASDs. The recovery takes the form of computing the unavailable block by a modulo 2 addition of the n−1 remaining blocks of a given parity group. Although DASDs in the array can suffer both hard as well as Byzantine faults, the worst case is to treat an array DASD as a hard failure and rewrite the data on the spare DASD, time permitting.

Referring now to FIG. **3**, there is shown the initial subsystem response to the detection of any anomaly in the read back or staging of data from a storage device, whether a single large device or as the logical equivalent formed from an array of small devices. More particularly, the IBM 3990 senses or detects an error or performance anomaly in step **301**. This occurs either by control logic **17, 19** polling any of the storage devices (FIG. **2**/**213, 41,** or **43**), any of the devices setting an interrupt in the control logic, or failure to respond. Relatedly, steps **303–315** ascertain whether the detected anomaly is of a type or nature for which the device long busy recovery procedure **315** should be invoked. Thus, if the anomaly is a hard device failure or a hard failure in the only path to a device as indicated in step **303**, then a long busy recovery process of step **315** will be invoked. Otherwise, as two or more paths to the device associated with the anomaly are operable, then resolution will be attempted without invoking step **315**. Clearly, steps **305, 307**, and **311** determine whether such multiple paths to a device are available. Parenthetically, for purposes of the method of this invention, a RAID 5 array is considered as a single device. As mentioned in the discussion of the embodiment in FIGS. **1** and **2**, a hierarchical subsystem of the IBM 3990/3390 type includes at least two failure-independent paths to each device. However, since paths may be unavailable for a variety of reasons on a permanent or intermittent basis, such a test is necessary for efficient subsystem use of fault management resources.

Referring now to FIG. **4**, there is shown the method of the invention after a determination that the error is not one correctable after retry and only one path to the device path is available. The correction may involve a variable duration recovery time. The recovery starts in step **315** with the presentation of a device long busy status signal to the host CPU **1, 3** by the control logic **17, 19**. In steps **401** and **403**. the control unit isolates the device by inhibiting two forms of access through suspension of read and write requests (step **401**) and pinning the destaging of tracks in NVS **31** for up to a maximum predetermined duration. In current practice, a maximum time interval/delay in the order of 90 seconds has proven effective. The control in step **405** passes to the device for invoking one or more recovery procedures for resolving the anomaly. It should also be appreciated that when a device operates in a recovery mode, it operates from a linear list of nested DRPs ordered on statistical assumptions as to causes of the anomaly.

The recovery procedures at the device level include correcting an error or erasure in a binary data stream of linear cyclic codewords using only nonzero syndromes, an erasure locator polynomial, and pointers. Also, the list of DRPs may include conditional branches to DRPs otherwise lower in list order such as where detection of a possible erasure or burst sets an interrupt in the device microprocessor.

While the invention has been described with respect to an illustrative embodiment thereof, it will be understood that various changes may be made in the method and means herein described without departing from the scope and teaching of the invention. Accordingly, the described embodiment is to be considered merely exemplary and the invention is not to be limited except as specified in the attached claims.

What is claimed is:

1. A method for detecting and correcting a defective operating state or condition of a hierarchical demand/responsive storage subsystem of the passive fault management type attaching a host CPU, said subsystem including a plurality of cyclic, tracked storage devices, an interrupt-driven, task-switched control logic, and means responsive to the control logic for forming at least one path of a set of paths coupling the host to at least one device, said host enqueuing one or more read and write requests to said subsystem, said subsystem control logic responsively interpreting each request and establishing a path to an addressed storage device, comprising the steps at the subsystem of:

(a) detecting an anomaly in the read back or staging of data from the device and executing a retry of the counterpart request by active or passive querying of said addressed device;

(b) in the event that the detected anomaly persists, presenting a long busy status signal to the host CPU by the

control logic, said long busy signal being an indication that the counterpart request has yet to be completed by the subsystem;

(c) inhibiting host access to the device by the control logic for no more than a predetermined time interval;

(d) ascertaining whether the inhibited device has returned to an operational state, and

(1) in the event the anomaly is resolved, setting an attention interrupt in the control logic by the device and terminating the device long busy signal in the host CPU by the control logic, and

(2) in the event that the time interval has been exceeded and the anomaly is not resolved, invoking one or more data recovery procedures including resetting the device by the control logic; and

(e) reporting status to the host CPU.

2. The method according to claim 1, wherein the step of ascertaining whether the inhibited device has returned to an operational state includes executing at least one step selected from the set of steps consisting of polling device status by the control logic, setting of an interrupt in the control logic by the device, and exceeding the predetermined (recovery) time.

3. The method according to claim 1, wherein the step of detecting an anomaly and retrying the request includes the steps of ascertaining whether at least two of the failure-independent paths to the device are operable and invoking step (b) where only one such path is ascertained as available.

4. The method according to claim 1, wherein the step of inhibiting access to the device includes the steps of suspending execution of any new read and write requests and pinning the destaging of any data to the device.

5. In a hierarchical demand/response storage subsystem of the passive fault management type, said subsystem being responsive to read and write requests from a host CPU for establishing access to at least one of a plurality of cyclic, multitracked storage devices over one path selected from a set of at least two failure-independent paths terminating in said device, said subsystem including means for detecting and correcting a defective operating state or condition in the subsystem or attached devices, whereby said detecting and correcting means further comprise:

means for detecting an anomaly in the read back or staging of a binary data stream from a device and for retrying said read back or staging;

means for ascertaining whether only one path to the device is operable, whether the anomaly persists after retry and, if so, for presenting a long busy status to the host CPU;

means for inhibiting host access to the device for up to a predetermined time interval;

means for terminating the long busy status in the host CPU responsive to an attention interrupt from the device indicative that the inhibited device has returned to an operational state and the anomaly has been resolved;

means responsive to the time interval having been exceeded and the nonresolution of the anomaly for invoking one or more data recovery procedures including resetting the device; and

means for reporting the current status of the device to the host.

6. An article of manufacture comprising a machine-readable memory having stored therein indicia of a plurality of processor-executable control program steps for detecting and correcting a defective operating state or condition of a hierarchical demand/responsive storage subsystem of the passive fault management type attaching a host CPU, said subsystem including a plurality of cyclic, tracked storage devices, an interrupt-driven, task-switched control logic, and means responsive to the control logic for forming at least one path of a set of paths coupling the host to at least one device, said host enqueuing one or more read and write requests to said subsystem, said subsystem control logic responsively interpreting each request and establishing a path to an addressed storage device, said plurality indicia of control program steps executable at the subsystem include:

(a) indicia of a control program step for detecting an anomaly in the read back or staging of data from the device and executing a retry of the counterpart request by active or passive querying of said addressed device;

(b) indicia of a control program step in the event that the detected anomaly persists for presenting a long busy status signal to the host CPU by the control logic, said long busy signal being an indication that the counterpart request has yet to be completed by the subsystem;

(c) indicia of a control program step for inhibiting host access to the device by the control logic for no more than a predetermined time interval;

(d) indicia of a control program step for ascertaining whether the inhibited device has returned to an operational state, and

(1) in the event the anomaly is resolved, for setting an attention interrupt in the control logic by the device and for terminating the device long busy signal in the host CPU by the control logic, and

(2) in the event that the time interval has been exceeded and the anomaly is not resolved, for invoking one or more data recovery procedures including resetting the device by the control logic; and

(e) indicia of a control program step for reporting status to the host CPU.

* * * * *